International Journal for Housing Science and Its Applications

Publish September 9, 2025. Volume 47, Issue 1 Pages 487–500

https://doi.org/10.70517/ijhsa47141

Innovative Research on English Oral Training Models in Digital Education Environments

Taotao Li[1,*] and Bao Chen[2]

[1] School of Foreign Languages, Tangshan Normal University, Tangshan, Hebei, 063009, China

[2] Technology Department, Tangshan Senpu Information Technology Co., Ltd., Tangshan, Hebei, 063000, China

Corresponding authors: (e-mail: litaotaoTSNU@126.com).

Abstract This paper addresses the need for English speaking training in a digital education environment by designing an intelligent English speaking training system based on deep learning. The system employs a semantic understanding model that integrates role information and historical dialogue context, utilizing a BERT-BiLSTM-CRF joint framework to achieve intent recognition and slot value filling. In the speech preprocessing stage, the system innovatively applies spectral entropy-based endpoint detection (VAD) to optimize the processing of low-energy speech signals, and combines pre-emphasis and Hamming window framing techniques to enhance recognition robustness. On the ATIS dataset, the system achieves an intent recognition accuracy of 99.19% and a slot filling accuracy of 97.24%, representing improvements of 0.7% and 1.1% over the best baseline, respectively. System performance testing shows that in real teaching interactions, the average response latency is 1024.2 ms, with 98% speech recognition accuracy, 98% task completion rate, and 92% pronunciation correction rate. In educational empirical studies, students' oral English scores significantly improved from 71.06 ± 15.99 points to 88.31 ± 8.54 points (+24.25%), the failure rate decreased from 24.51% to 0%, and the excellent rate (>90 points) increased from 16.67% to 48.04%. The learning attitude questionnaire showed that the number of students who fully agreed with "fluent English speaking" increased from 41 to 80 (+95.1%), and the willingness to persist in daily training increased by 252.4% (from 21 to 74 students). The study indicates that the system effectively enhances oral training efficiency through deep semantic understanding and multimodal interaction design, providing technical support for digital English teaching.

Index Terms English speaking, deep learning, BERT-BiLSTM-CRF, historical influence vector, speech signal processing

I. Introduction

In the context of the digital age, cross-cultural communication across languages has become increasingly frequent and in-depth. The social context now places higher demands on individuals' English core competencies and comprehensive abilities, particularly in terms of innovative thinking skills and oral communication abilities, which require greater emphasis [1]-[4]. Oral communication training, as a new teaching model, can alleviate students' learning stress and help them develop personal learning confidence during the training process [5], [6]. Students can utilize oral communication training to promote their physical and mental health development and experience the joy of English learning during the training process, thereby genuinely enhancing their overall English proficiency [7]-[9]. Teachers, while maintaining students' basic learning abilities, can use oral communication training to create a relaxed and open learning environment, thereby stimulating students' interest in English learning [10]-[12]. Therefore, teachers need to innovate oral communication training to achieve certain teaching effects, thereby enabling students' overall English proficiency to develop comprehensively.

Under digital technology, university English oral communication training models exhibit new characteristics such as knowledge cloudification, content aggregation, and audience diversification. The deep development of digital technology has greatly activated new momentum in higher education [13]. In terms of personalization, digital English oral communication training demonstrates significant advantages and potential, capable of providing customized learning plans based on learners' individual differences and needs [14], [15]. In terms of intelligence, the application of technologies such as natural language processing, machine learning, and speech recognition in digital English speaking training enables the training system to automatically identify learners' pronunciation, intonation, speaking speed, and other speaking elements, providing immediate feedback and correction suggestions [16]-[17]. In summary, under the backdrop of educational digital transformation, how to build a digital education development ecosystem and reconstruct new teaching models for English speaking training is the core issue of current curriculum reform and teaching innovation.

At present, some online learning platforms have become important channels for students to train their English speaking skills. These platforms provide learners with fixed templates and language patterns for college students to imitate and practice,

which is beneficial for improving their speaking abilities. Literature [18] applied the Moodle Language Management System (LMS) to university English speaking classrooms, finding that the digital tools it contains and the digital learning environment it creates effectively increase students' learning motivation, help improve their speaking skills, and reduce anxiety during spoken communication. Literature [19] introduces an IoT-based online English speaking teaching platform that creates a virtual teaching environment conducive to students' speaking expression. Combined with a speaking training correction system, this further enhances the effectiveness of the online speaking teaching platform. Literature [20] indicates that online learning platforms can provide students with interactive tools, real-time feedback, and communication opportunities. When combined with English speaking instruction, they play an important role in improving learners' speaking ability, pronunciation, and fluency. Literature [21] constructs an English speaking mobile teaching platform integrating virtual reality technology. This platform demonstrates rapid response and convenient application advantages at the student, teacher, and administrator levels, achieving good interactive practice effects. Literature [22] emphasizes that video technology is an important auxiliary tool for enhancing learners' oral skills, proposing the use of the FlipGrid video discussion platform as a new platform for training learners' oral communication abilities and presentation skills. It has played a significant role in training learners' oral communication skills in scientific fields. Literature [23] clarifies that oral communication in a network environment has characteristics such as real-time, interactivity, and personalization. On one hand, network platforms provide learners with a wealth of learning resources; on the other hand, they create oral communication scenarios, providing strong support for language learning development. Literature [24] discusses the positive effects of social media platforms on enhancing learners' English oral communication skills. Rich and interesting video content not only promotes the acquisition of oral communication knowledge but also enhances learners' desire to use practical English for communication. However, while digital learning platforms provide learners with abundant opportunities for oral communication training, they also reduce their ability to identify and judge the accuracy of information, which is detrimental to learners' ability to identify and promptly correct errors in their learning.

Meanwhile, intelligent speech recognition software and other digital tools for digital English speaking training also provide different auxiliary effects for speaking instruction, to some extent enhancing students' creativity and making their speaking expressions more natural and fluent. Literature [25] developed an adaptive learning system suitable for English speaking instruction, incorporating machine learning-based knowledge representation methods and speech recognition functions to create an educational environment tailored to learners' individual learning styles. Literature [26] argues that a good oral teaching environment is the key to promoting reforms in English oral teaching. The proposed intelligent oral dialogue system evaluates pronunciation quality and provides corrective feedback, offering accurate guidance for English oral training in higher education institutions. Literature [27] established a speech recognition model based on long short-term memory (LSTM) neural networks, which can analyze and recognize multiple parameters in speech data, providing assistance for oral pronunciation training. Literature [28] addresses the issues of inadequate ability assessment and low feedback latency in traditional oral teaching processes by proposing an English oral data analysis and feedback system based on support vector machines (SVM), offering an innovative solution to enhance learners' personalized learning levels and self-directed learning abilities. Literature [29] points out that the application of natural language processing (NLP) and speech recognition technology can provide real-time interactive experiences and personalized feedback for English oral teaching, significantly enhancing the effectiveness of learners' English oral training. Literature [30] proposes a dual-phonation model to optimize pronunciation effects in English oral training. By employing visual weighting functions and dynamic feedback mechanisms during the pronunciation process to optimize mouth shape parameters, it effectively addresses the generalization issues in traditional speech recognition methods, thereby enhancing the reliability of intelligent oral training. It is evident that the integration of digital technology enhances learners' initiative and creativity in English speaking training. However, existing technologies have limitations in handling interactive learning environments and language-cultural identity, which need to be addressed to further improve speaking training outcomes.

This study focuses on innovating English speaking training models in digital educational environments, with the core objective of designing and implementing an intelligent English speaking training system based on deep learning. The system constructs a pipeline processing framework comprising four modules: natural language understanding, dialogue state tracking, dialogue strategy, and natural language generation. It employs a deep semantic understanding model that integrates role information and historical dialogue context to accurately parse user spoken input, initializes the BiLSTM using historical influence vectors, and combines current sentences for intent recognition and slot value filling. Finally, it presents a loss function for jointly training intent recognition and slot value filling tasks to optimize model performance. It is supplemented by necessary speech signal preprocessing techniques to ensure input quality, ultimately achieving natural and fluent human-machine spoken interaction. The speech signal preprocessing includes three steps: pre-emphasis uses filters to compensate for the attenuation of high-frequency speech signals, enhancing the high-frequency resolution of the signal. Framing and windowing divides continuous non-stationary speech signals into short-term stationary frames, and applies a Hamming window for windowing processing to meet the requirements of short-term analysis. Spectral entropy-based endpoint detection identifies the valid start and end points of speech (VAD) to exclude silent and noisy segments. This method distinguishes valid

speech from background noise by calculating the short-term spectral entropy and its distribution characteristics of speech frames, particularly suitable for speech signals with low energy.

II. Design of an English speaking training system based on deep learning

II. A. Intelligent dialogue design for English speaking training systems

When designing intelligent dialogues, it is important to consider that students need to engage in conversations within specific scenarios and practice their English speaking skills during the process. Therefore, this paper designs task-based dialogues based on hotel scenarios to help users engage in conversations or complete specific tasks within the scenario based on specific themes. In the specific implementation of task-oriented dialogues, the entire dialogue process is handled using a pipeline approach. The intelligent dialogue module first identifies and understands the user's input, representing it in a form that the computer system can understand. It then responds to the user based on the current dialogue turn, state information, and predefined dialogue strategies. Finally, it converts these execution results into natural language that humans can understand and feeds them back to the user.

In the system design process, the intelligent dialogue structure based on the pipeline method is shown in Figure 1, which primarily includes four key modules: natural language understanding, dialogue state tracking, dialogue strategy, and natural language generation.

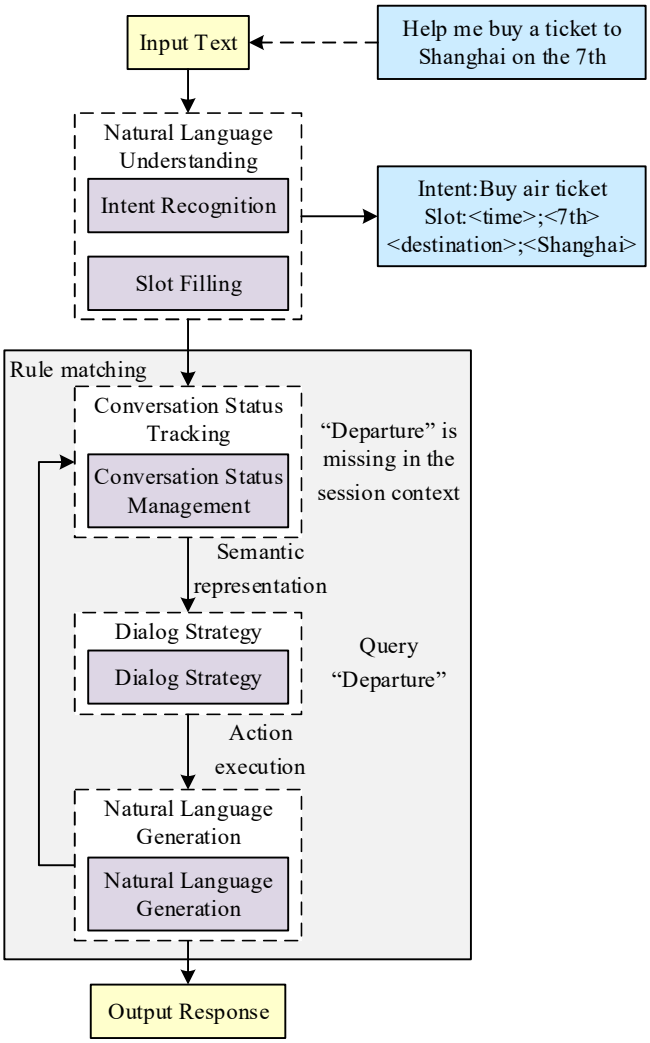

Figure 1: Intelligent Dialogue Module

II. B. Spoken language understanding method based on role information and historical influence vectors

The core foundation of building an efficient intelligent dialogue system lies in accurately understanding the semantic meaning of users' spoken input (NLU). Therefore, in order to significantly improve the system's accuracy in identifying user intent and

key information, especially in multi-round, role-based dialogue scenarios, we have innovatively proposed a deep semantic understanding model that integrates role information and historical dialogue context (historical influence vectors). The specific design is as follows.

II. B. 1) Semantic understanding model based on role information

First, the sentence undergoes BERT preprocessing to obtain the sentence's word vectors. Next, historical influence vectors are extracted by calculating the influence weights of historical sentences using a multi-layer perceptron (MLP). The sentences are grouped based on the roles in the multi-round dialogue, and the historical influence vectors are obtained by combining the grouped sentences with the influence weights calculated by the MLP. Finally, this vector is used as the initial value for the head and tail node hidden layers of the BiLSTM. The output of each hidden layer state and the tail node hidden layer state is obtained, and the sentence intent probability information is obtained through the Softmax function. The remaining outputs of the bidirectional LSTM model are passed to the CRF layer, represented using the BIO method, to obtain the sentence slot value probability distribution, ultimately yielding the sentence intent and slot value probability distribution. The semantic understanding model adopted in this paper is shown in Figure 2.

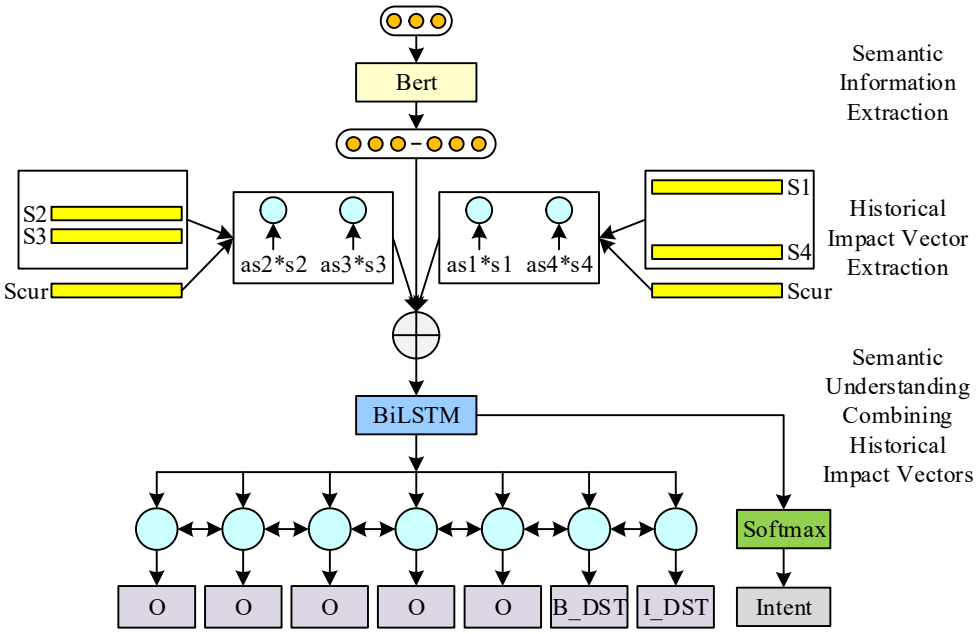

Figure 2: Semantic understanding model

II. B. 2) Semantic Information Extraction

Semantic information extraction uses the BERT model. This model consists of 12 layers of Transformer architecture, with each Transformer having a hidden layer size of 768, a total of 108,863,251 parameters, and 12 self-attention heads. By combining the pre-trained BERT model with a downstream-specific model, fine-tuning is performed to train the model, calculate the total loss value, compute gradients, and train the network through backpropagation.

The model input is $L_0 = \{(S_i, y_i^i, y_i^s)\}, i \in [1, n]$, where S_i represents the dialogue sentence sample, n is the total number of samples, y_i^i is the label of the intent for sample i, and y_i^s is the slot value entity label for sample i. The result of sentence segmentation is $S_i = (c_1, c_2, c_3, ...c_T)$. After inputting the sentence into the BERT model, the embedding vectors of each word are obtained, which serve as the semantic representation of the sentence $S = [h_1, h_2, ...h_t]$.

II. B. 3) Semantic understanding combined with historical influence vectors

To further improve semantic understanding accuracy, we propose an English spoken language semantic understanding method based on a BiLSTM network, which is built upon role-based historical influence vectors. This method enables intent recognition and slot value filling for the current sentence. The historical sentence influence vector V_{hif} is used as the initial value for the head and tail node hidden layers of the BiLSTM network. The current sentence semantic representation

$S_{cur} = [S_1, S_2, \cdots, S_n]$ is used as the model input, thereby outputting the states of each hidden layer and the tail node hidden layer. The semantic understanding model incorporating role information is shown in Figure 3.

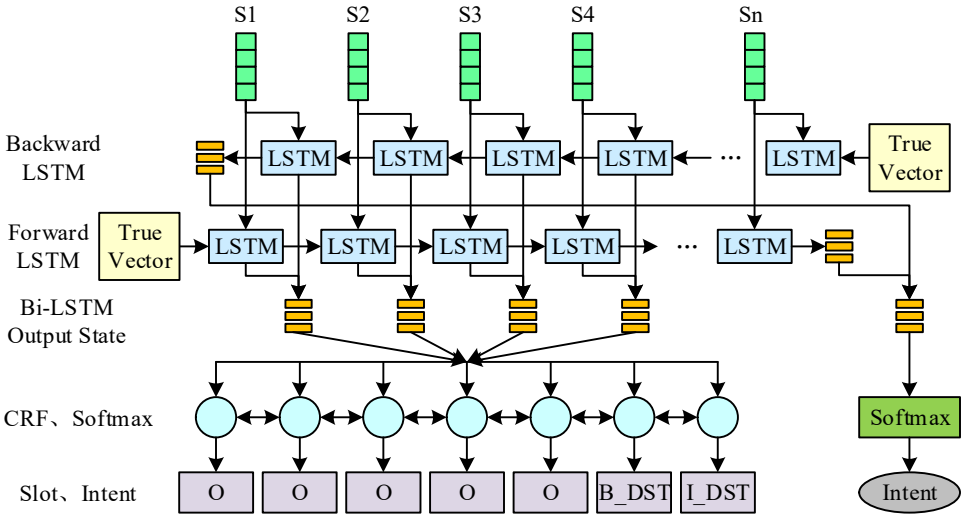

Figure 3: Semantic understanding combined with historical influence vectors

As shown in Figure 3, semantic understanding combined with historical influence vectors mainly consists of three steps, specifically:

(1) Calculate the intent classification.

Use the Softmax function to perform intent classification prediction on the head and tail nodes of the BiLSTM network. The specific calculation formula is:

$$V_{cur} = BiLSTM\left(S_{cur}, V_{hif} \times W_{hif}\right) \tag{1}$$

$$Intent = Soft\max\left(V_{cur} \times W_{intent}\right) \tag{2}$$

In the above equation, W_{intent} denotes the weight matrix of the fully connected layer for intent classification; V_{hif} denotes the historical influence vector containing role information; W_{hif} denotes the historical sentence weight matrix; V_{cur} denotes the vector obtained by concatenating the output vectors of the hidden layers at the beginning and end nodes of the BiLSTM network. After mapping V_{cur} to the interval $[0, 1]$ via the Softmax function, it is used as the predicted probability value, and the classification with the highest probability value is selected as the model's intent prediction result.

(2) Calculate slot value classification.

After concatenating the nodes before and after the sentence, CRF is used to classify and predict the slot values of the input sentence. The specific expression is:

$$Slot(V_i) = CRF\left(Soft\max(h_i)\right) \tag{3}$$

In equation (3), h_i represents the LSTM hidden layer output vector corresponding to word i. After using the Softmax function to distinguish between intent categories, the hidden layers of the BiLSTM network can be output, and the slot value classification can be obtained through CRF.

The CRF model can judge the correctness of input sentences based on the logical relationships of the language, and then train and learn according to sequence annotation rules to obtain the best recognition results. The expression for the CRF model is:

$$Score(l \mid s) = \sum_{j=1}^{m} \sum_{i=1}^{n} \lambda_j f_j\left(s, i, l_i, l_{i-1}\right) \tag{4}$$

In equation (4), $f_j(s, i, l_i, l_{i-1})$ represents the feature function, which calculates the score value of the i th word in sentence s; l_i and l_{i-1} represent the annotated part-of-speech tags of the i th word and the $(i-1)$ th word, respectively. The scores of the n words are calculated using the results of the m feature functions and then summed to obtain the score value. This score value is then exponentiated using Equation (5) to obtain the annotation probability value $p(l \mid s)$ for the current sentence. The higher the probability value, the higher the reliability of the annotation result for the sentence.

$$p(l \mid s) = \frac{\exp[score(l \mid s)]}{\sum_{l'} \exp[score(l \mid s)]} = \frac{\exp\left[\sum_{j=1}^{m}\sum_{i=1}^{n} \lambda_j f_j(s, i, l_i, l_{i-1})\right]}{\sum_{l'} \exp\left[\sum_{j=1}^{m}\sum_{i=1}^{n} \lambda_j f_j(s, i, l_i, l_{i-1})\right]} \tag{5}$$

(3) Calculate the loss function for slot value filling and intent recognition.

The loss function formula can be expressed as:

$$\tau_{slot} = -\sum_{j=1}^{n}\sum_{i=1}^{m} \hat{y}_{slot}^{i} \log\left(y_{slot}^{i}\right)$$

$$\tau_{\text{int}ent} = -\sum_{i=1}^{k} \hat{y}_{\text{int}ent}^{i} \log\left(y_{\text{int}ent}^{i}\right) \tag{6}$$

In the above equation, k denotes the number of intent categories in the dataset; n denotes the number of words after tokenization; m denotes the number of slot value categories; $y_{\text{int}ent}^{i}$ and $\hat{y}_{\text{int}ent}^{i}$ denote the true intent and predicted intent, respectively; $\hat{y}_{slot}^{i}$ and y_{slot}^{i} denote the true slot value and predicted slot value, respectively.

II. C. Speech Signal Preprocessing

The aforementioned semantic understanding module primarily processes text-based input. However, in real-world speech interaction systems, user input is initially presented in the form of speech signals. To ensure that the subsequent speech recognition module can accurately convert speech into text and ultimately feed it into the semantic understanding module, effective preprocessing of the raw speech signals to enhance their quality is of critical importance.

The purpose of speech signal preprocessing is to optimize the frequency bands of speech information, ensuring a smooth and stable speech signal spectrum while improving speech resolution in both high and low frequency bands. Preprocessing primarily includes three components: pre-emphasis, frame division and windowing, and endpoint detection.

II. C. 1) Preloading

High-frequency signals in voice messages are prone to significant attenuation, and the rate of attenuation is relatively fast, which can result in incomplete voice signals before they are put into use. To address this issue, voice signals are pre-emphasized. This process essentially involves filtering, which can be expressed as:

$$H(z) = 1 - \alpha z^{-1}, 0.9 \le \alpha \le 1.0 \tag{7}$$

In the formula, α is the pre-weighting coefficient, with a value close to 1.

II. C. 2) Frame splitting and windowing

The processing of time-varying signals is quite challenging, and speech signals are a type of unstable time-varying signal. However, due to the unique nature of human speech information (produced by the human vocal tract and stable within a time range of 10 ms to 30 ms), short-time analysis methods can be used for processing. The prerequisite for speech signal processing is to segment the complete speech information into multiple stationary segments for subsequent processing, a process commonly referred to as frame division. The prerequisite for frame segmentation is to ensure the integrity of the speech information. Essentially, frame segmentation of speech signals is a form of windowing processing, which can be expressed as:

$$x_n(m) = x_n(m) * \omega(m) \tag{8}$$

There are many types of windowing functions corresponding to windowing processing. Functions used for frame processing include: Hamming window, Hanning window, and rectangular window. Since this study requires spectrum analysis, the Hamming window with the smallest spectrum leakage is selected. The corresponding formula is:

$$\omega(n) = \begin{cases} 0.54 - 0.46\cos\left[2\pi n / (N-1)\right] & 0 \le n \le N-1 \\ 0 & Other \end{cases} \tag{9}$$

In the formula, N is the frame length.

II. C. 3) Spectrum entropy endpoint detection

The acquired speech signals contain a significant amount of noise, including silent segments and noise segments. This unnecessary information increases the computational load of data processing and also affects the quality of speech signal

processing. Endpoint detection is one of the effective methods to address this issue. The principle of this method is to locate the positions of useful information and then remove the unnecessary information segments. In endpoint detection operations, the most commonly used algorithms are the zero-crossing rate double threshold method and the short-term energy algorithm. However, since the speech signals targeted in this study have relatively low signal energy, the short-term energy algorithm cannot be selected. Therefore, this study adopts the spectral entropy method for endpoint detection.

Endpoint detection using spectral entropy First, any frame of speech signal data to be denoised is subjected to FFT transformation to obtain the corresponding frequency component spectral line energy spectrum $Y_i(k)$, from which the corresponding frequency component probability density function can be obtained:

$$p_i(k) = \frac{Y_i(k)}{\sum_{l=0}^{N/2} Y_i(l)} \tag{10}$$

In the formula, N is the FFT length.

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

$$H_i = -\sum_{k=0}^{N/2} p_i(k) \log p_i(k) \tag{11}$$

The following relationships exist:

$$H(P) = H(p_1, p_2, \cdots, p_q) = H(1/q, 1/q, \cdots, 1/q) = \log q \tag{12}$$

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

To further enhance the performance of the spectral entropy method, reasonable settings are made:

(1) To enable clearer distinction between valid and invalid speech segments during endpoint detection, reasonable settings are made. The frequency range is set to 250 Hz to 3400 Hz. If the frequency of any spectral line is f_k, then the corresponding values are:

$$Y_i(k) = 0\left(f_k < 300\ Hz\ or\ f_k > 250\ Hz\right) \tag{13}$$

(2) Due to the low spectral entropy values corresponding to some special noises, it is impossible to remove these noise signals when using the spectral entropy method for noise reduction. This problem can be solved by setting the normalized spectral probability density, i.e., setting the density upper limit to:

$$p_i(k) = 0 \quad p_i(k) > 0.9 \tag{14}$$

III. Performance verification of a deep learning-driven English speaking training system

A spoken language understanding model based on role information and historical influence vectors, as well as speech preprocessing technology, has been developed. To validate the actual effectiveness of this system, Chapter 3 will conduct empirical evaluations from two dimensions: technical indicators and teaching applicability.

III. A. Research on spoken language understanding technology based on deep learning

III. A. 1) Dataset and experimental setup

To demonstrate the effectiveness of the proposed spoken language understanding method based on role information and historical influence vectors, this chapter conducts experiments on two public English dialogue datasets, ATIS and SNIPS. ATIS is an aviation-related dataset widely used in spoken language understanding tasks. SNIPS is a dataset collected from the Sinps voice assistant, whose dialogues cover multiple domains with a relatively balanced category distribution. The training samples for the two datasets are 5,278 and 17,239, respectively, with the validation and test sets divided in a 7:1:2 ratio.

This chapter evaluates the experimental results using three evaluation metrics. For intent recognition and slot filling tasks, this chapter uses accuracy and F1 score for evaluation, where F1 score is a comprehensive consideration of precision and recall.

This chapter uses the BERT pre-trained model as the encoding layer. BERT was pre-trained on two corpora: Bookcorpus (1 billion words) and English Wikipedia (3 billion words). The learning rate is pre-set to 5e-4 and is adaptively adjusted based on the loss function of the development set.

III. A. 2) Analysis of comparative experimental results

The experiment primarily compares the proposed method with Joint Seq, Attention Bi RNN, Slot-Gated Full Atten, Slot-Gated Intent Atten, Self-Attentive Model, Bi-Model, CAPSULE-NLU, SF-ID Network, Stack-propagation, BERT-baseline,

Stack+BERT, and the proposed method based on role information combined with historical influence vectors for spoken language understanding. Tables 1 and 2 present the performance comparisons of each model on the ATIS and SNIPS datasets, respectively.

Table 1: Performance comparison of each model on the ATIS dataset

	Intent		Slot		Sentence	
	Accuracy	F1	Accuracy	F1	Accuracy	F1
Joint Seq	96.53	95.43	87.17	86.47	73.68	72.71
Attention Bi RNN	97.22	96.31	87.63	87.06	74.41	73.68
Slot-Gated Full Atten	97.96	96.28	88.44	87.82	76.55	75.84
Slot-Gated Intent Atten	97.09	95.20	89.09	88.86	77.07	76.47
Self-Attentive Model	97.02	96.28	89.92	89.35	80.81	80.34
Bi-Model	97.17	95.51	90.27	89.62	83.97	82.90
CAPSULE-NLU	96.94	95.26	91.02	90.4	79.45	78.63
SF-ID Network	97.38	96.76	90.58	90.29	80.23	79.51
Stack-propagation	98.10	96.41	92.92	92.50	87.33	86.23
BERT-baseline	98.44	97.31	95.49	95.25	89.09	88.45
Stack+BERT	98.49	97.20	96.14	95.57	93.49	92.43
OURS	99.19	97.65	97.24	96.71	94.52	93.87

On the ATIS dataset, our method achieves a significant lead over all baseline models in the intent recognition task, with an accuracy of 99.19% and an F1 score of 97.65%. Compared to the current state-of-the-art Stack+BERT model (accuracy of 98.49% and F1 score of 97.20%), our method improves accuracy by 0.7 percentage points. Traditional models such as Slot-Gated Full Atten (97.96%) and BERT-baseline (98.44%) are both outperformed. In the slot filling task, the proposed method ranks first with an accuracy of 97.24% and an F1 score of 96.71%, outperforming the second-place Stack+BERT (96.14% accuracy) by 1.1 percentage points. The proposed method achieves the best sentence-level accuracy (94.52%) and F1 score (93.87%), outperforming Stack+BERT (93.49% accuracy) by 1.03 percentage points. Early models such as Joint Seq (73.68%) lag behind the proposed method by over 20 percentage points, demonstrating the superiority of language understanding methods that combine role information and historical influence vectors.

Table 2: Performance comparison of each model on the SNIPS dataset

	Intent		Slot		Sentence	
	Accuracy	F1	Accuracy	F1	Accuracy	F1
Joint Seq	92.41	91.07	94.01	94.31	80.07	79.43
Attention Bi RNN	91.80	90.86	94.88	95.13	78.69	77.67
Slot-Gated Full Atten	93.83	92.53	95.65	96.09	82.24	81.07
Slot-Gated Intent Atten	94.44	93.27	94.62	95.16	81.53	80.47
Self-Attentive Model	95.59	94.29	94.43	95.08	84.50	83.74
Bi-Model	96.30	94.96	95.11	95.51	83.34	82.39
CAPSULE-NLU	95.45	94.39	93.91	94.59	85.43	84.27
SF-ID Network	96.35	94.93	96.80	97.32	86.06	85.24
Stack-propagation	96.83	95.82	96.65	97.30	86.30	85.13
BERT-baseline	97.31	96.02	97.28	97.71	88.08	87.26
Stack+BERT	97.84	97.08	97.68	97.99	88.75	87.75
OURS	98.38	97.18	98.19	98.49	89.67	88.73

In the SNIPS dataset, our method achieved a leading accuracy rate of 98.38% and an F1 score of 97.18% in the intent recognition task, surpassing the 97.84% accuracy rate of Stack+BERT by 0.54 percentage points. In the slot filling task, the proposed method achieves a record-breaking accuracy of 98.19% and an F1 score of 98.49%, surpassing the 97.68% accuracy of Stack+BERT by 0.51 percentage points. The slot filling performance of SF-ID Network (96.80%) and Bi-Model (95.11%) lags significantly behind the proposed method, with the largest gap being 3.08 percentage points. In sentence-level tasks, the proposed method leads with an accuracy of 89.67% and an F1 score of 88.73%, but the advantage over the ATIS dataset has narrowed, with only a 0.92 percentage point lead over Stack+BERT. Under multi-domain data, the gaps between various models and the proposed method indicate that historical context integration is more challenging in complex scenarios.

III. A. 3) Ablation experiment

To demonstrate the effectiveness of each module in the spoken language understanding method based on role information and historical influence vectors, this chapter uses a BERT-based joint model as the baseline model, adding independent modules to it one by one, and analyzes the results through comparative experiments. Additionally, different feature fusion methods are employed to compare their impact on model performance. This chapter selects four commonly used feature fusion methods for experimentation: Mean Pooling, Max Pooling, Concatenate, and Attention. The ablation experiment results on the ATIS dataset are shown in Table 3.

Table 3: The results of the ablation experiment

	Intent		Slot		Sentence	
	Accuracy	F1	Accuracy	F1	Accuracy	F1
BERT-baseline	98.44	97.31	95.49	95.25	89.09	88.45
BERT+Role	98.59	97.42	95.98	96.67	91.33	89.05
BERT+Hist	98.89	97.53	96.35	96.55	92.05	89.59
FULL	99.19	97.65	97.24	96.71	94.52	93.87
By Mean Pooling	98.53	97.34	95.09	95.64	92.39	91.83
By Max Pooling	98.67	97.38	96.06	96.63	91.93	91.51
By Concatenate	98.93	97.57	96.76	97.15	93.91	93.42
By Attention	99.19	97.65	97.24	96.71	94.52	93.87

It can be seen that after adding the role module (BERT+Role), the sentence-level accuracy improved from the baseline of 89.09% to 91.33% (+2.24%), indicating that role grouping effectively distinguishes between user and system discourse preferences. The slot filling F1 value improved by 1.42% (95.25% → 96.67%), indicating that role information assists entity parsing. Adding the history module (BERT+Hist) significantly improved slot filling accuracy to 96.35%, compared to the baseline model's accuracy of 95.49%, an increase of 0.86%, validating the critical role of historical context in entity consistency. The sentence-level F1 score improved by 1.14% (88.45% → 89.59%), reflecting the importance of temporal information in modeling dialogue state.

The complete model achieves the best performance in terms of intent recognition accuracy (99.19%) and sentence-level F1 score (93.87%), with a 0.6% improvement in intent accuracy compared to the BERT+Role model that only adds the role module. and a 4.82% improvement in sentence F1 score, which is a 2.47% improvement in sentence accuracy compared to BERT+Hist (92.05% → 94.52%), indicating that the role and history modules complement and enhance each other, with role information optimizing discourse subject perception and history vectors reinforcing cross-turn semantic dependencies.

The attention mechanism is the optimal fusion scheme for role and history vectors, with its dynamic feature selection capability adapting to the dynamic nature of dialogue. Mean pooling performs the weakest in intent recognition (98.53%) and slot filling (95.64% F1), as homogenization processing weakens key features. Max Pooling is slightly better than Mean Pooling, with a slot filling F1 of 96.63%, but the sentence-level accuracy (91.93%) is still 4.59% lower than the attention mechanism. Vector concatenation achieved an accuracy of 93.91% in sentence-level tasks, close to the attention mechanism's 94.52%, indicating that simple concatenation can preserve multimodal features but has lower computational efficiency. The attention mechanism feature fusion method achieves the highest intent recognition accuracy of 99.19%, which is 0.26% higher than vector concatenation, with a slot filling F1 score of 96.71%, confirming that attention can dynamically weight role and historical features.

III. A. 4) Case Study

To more accurately illustrate the role of the semantic understanding method based on role information and historical influence vectors under the attention mechanism, this chapter compares the method proposed in this paper with the BERT baseline model. Two sentences with the same intent and slot positions but different slot words are predicted by the network model, and the attention distribution matrix of the last layer of the Transformer structure is used to draw a heatmap for comparison. The similarity analysis of sentences for BERT and the proposed model is shown in Figures 4, 5, and 6.

Sentence 1 is *"I want to book a double room in Paris from June 10th to 15th."*, Sentence 2 *"I want to order a single room in Tokyo for July 5th to 11th."*

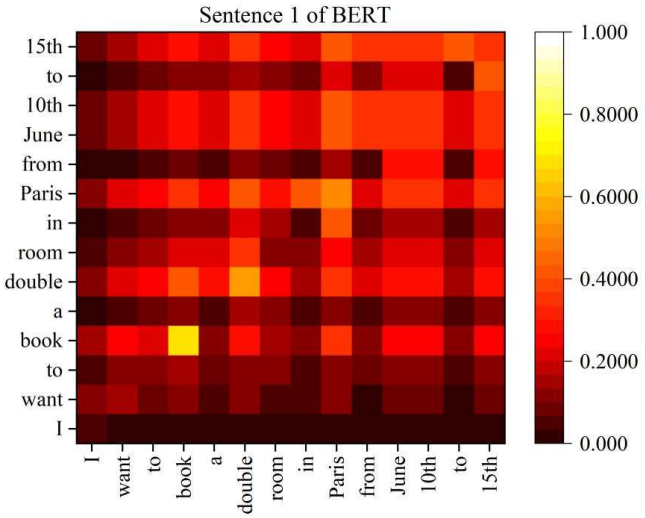

Figure 4: Thermal analysis diagram of sentence 1 under BERT

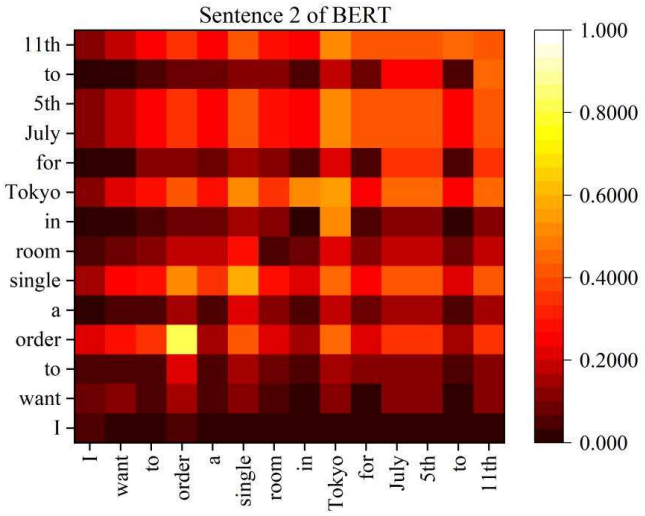

Figure 5: Thermal analysis diagram of sentence 2 under BERT

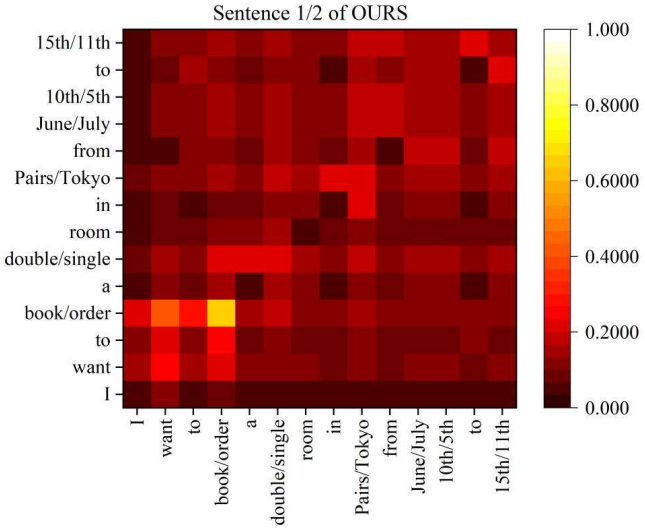

Figure 6: Thermal analysis diagram of sentence 1/2 under OURS

As can be seen from the figure, although the two sentences express the same intent of "booking a hotel," the slot positions 'destination' and "date" are not the same. Due to the specific differences in the slot words, the baseline model does not distribute attention equally between the two slot words. In the baseline model, "date" is treated as one of the important features that may appear in the intent "book a hotel." However, in the ATIS dataset, many examples of different intents include the slot "booking date," which demonstrates that the implicit joint model selects some incorrect slot information as classification features for the intent recognition task, thereby affecting the model's performance. Due to verb differences (book and order) and slot value changes (double/Paris and single/Tokyo), the baseline model BERT correctly identified Sentence 1 as book_hotel with a confidence of 92%; Sentence 2 was misclassified as order_food with a confidence of 78%, as "order" is more frequently associated with dining intents in the training set.

Figure 6 shows the heat distribution map of the model in this paper for these two similar sentences. In the method proposed in this chapter, several slot-corresponding words obtain similar weights in the attention distribution, which facilitates the model to learn the patterns between similar sentence structures and avoid being affected by the specificity of slot words. Both sentences are identified as user requests, and system response-type features are automatically filtered out. The historical vector reinforces the importance of the [city] slot and weakens the importance of date details. Additionally, the method proposed in this paper does not use "destination" or 'date' as features for the "book a hotel" intent, demonstrating that the semantic understanding method based on role information and historical influence vectors under the attention mechanism can help the model select more effective features and enhance its ability to extract deep semantic information from sentences.

III. B. Performance testing of an English speaking training system based on deep learning

To further evaluate the system's overall performance in real-world voice scenarios through human-machine interaction testing, we primarily employed human evaluation methods for performance assessment, focusing on key metrics such as response latency and speech recognition accuracy—critical indicators for educational applications. Specifically, five volunteers interacted with the system, with each volunteer teaching 10 objects. Following the teaching session, the volunteers posed questions to the system based on the content they had taught. The evaluation was conducted in two parts: teaching and questioning. The teaching process primarily focused on whether the system could complete the learning of several specific attributes of the object smoothly and comprehensively. The questioning part primarily focused on whether the system could apply the knowledge it had learned to correctly answer the user's questions. The evaluation results of the system are shown in Table 4 below.

Table 4: The evaluation results of the system

Evaluation items	Volunteer 1	Volunteer 2	Volunteer 3	Volunteer 4	Volunteer 5	Average
Average conversation length	6.80	6.46	7.33	6.63	6.81	6.81
Average number of conversation rounds	7.3	10.8	8.1	5.2	7.1	7.7
Response delay	1204ms	947ms	1034ms	834ms	1102ms	1024.2ms
Knowledge acquisition rate	100%	100%	100%	100%	90%	98%
Speech recognition accuracy rate	100%	90%	100%	100%	100%	98%
Pronunciation correction rate	100%	80%	90%	90%	100%	92%
Task completion rate	100%	100%	100%	100%	90%	98%
Inductive questioning rate	100%	90%	100%	100%	100%	98%

The system performed exceptionally well in multi-dimensional testing by five evaluators. Interaction efficiency: The average conversation length was 6.81 rounds, and the average number of conversation rounds was 7.7, indicating that the system can efficiently complete target conversations. The average response latency was 1024.2ms, ranging from 834ms to 1204ms. In terms of core functionality performance, the knowledge acquisition rate and task completion rate both reached 98%, with only evaluator 5 experiencing a 10% deficiency in knowledge acquisition. Speech recognition accuracy was 98%, with only evaluator 2 experiencing a 10% error rate. In terms of teaching assistance capabilities, pronunciation correction rate was 92%, with evaluator 2's 80% being the lowest value, and summary question rate was 98%, indicating that the system can effectively assist in oral training, but the stability of pronunciation correction needs further improvement.

IV. Research on the teaching results of English speaking training systems

Chapter 3 verifies the technical feasibility and basic performance of the deep learning-based English speaking training system. To further explore its application effectiveness in actual teaching, a teaching study was conducted over one semester with 102 students majoring in English education at a certain university in 2024 as the research subjects. The English speaking training system designed in this article was applied. The educational effectiveness of the system was analyzed through questionnaire surveys and performance tests.

IV. A. Questionnaire Survey Analysis

After a semester of teaching experiments, relevant questionnaires were distributed to assess students' attitudes toward English speaking, classroom performance, classroom expectations, after-class extension, and speaking improvement. A total of 73 questionnaires were distributed and all were returned, with a 100% return rate.

IV. A. 1) Analysis of Attitudes Toward Learning Spoken English

This study examined attitudes toward English speaking learning from five perspectives: A1: Desiring to speak English fluently; A2: Desiring to communicate effectively with people from English-speaking countries; A3: Being interested in the culture of English-speaking countries and desiring to understand English culture; A4: Believing that English speaking learning is important and having confidence in mastering speaking skills; A5: Being willing to spend time daily on English speaking practice. The results of the pre- and post-survey comparison of students' attitudes toward English speaking learning are shown in Table 5.

Table 5: The comparison results of one's attitude towards learning spoken English

		Exactly Consistent	Consistent	Generally Consistent	Inconsistent	Definitely Inconsistent
A1	Before	41	20	26	10	5
	After	80	14	5	2	1
A2	Before	35	37	8	15	7
	After	76	19	3	3	1
A3	Before	39	27	20	11	5
	After	81	16	3	2	0
A4	Before	27	28	21	17	9
	After	69	21	6	4	2
A5	Before	21	28	32	14	7
	After	74	15	7	4	2

The survey questionnaire shows that the teaching experiment significantly improved students' enthusiasm for learning English speaking skills and strengthened their willingness to learn. The number of students who "strongly agree" that they want to "speak English fluently" increased from 41 to 80, a 38.24% increase, while the number of students who "strongly disagree" decreased from 15 to 3. The number of people who "completely agree" that they "train daily" increased by 253% (from 21 to 74). The percentage of people who "completely agree" that they are interested in English culture rose from 39 to 81, accounting for 79.41% of the total. The number of people who "completely agree" that spoken English is important and that they are confident they can learn it well increased from 27 to 69, while the number of those who disagree decreased by 19.61% (from 26 to 6).

IV. A. 2) Analysis of Expectations in English Speaking Classes

The questionnaire items regarding students' expectations for oral English classroom activities prior to the teaching experiment were as follows: EB1: Hope for a more relaxed and lively classroom atmosphere; EB2: Hope for more teaching aids such as images, music, audio, video, and physical teaching materials in the classroom; EB3: Hope for more diverse classroom activities. After the teaching experiment, the questionnaire on expectations for the English oral communication training system in the classroom included the following questions: EA1: The classroom atmosphere under the English oral communication training system is relaxed; EA2: The system provides more teaching aids such as images, music, audio, video, and physical teaching materials for classroom instruction; EA3: Classroom activities are relatively diverse. The results of the pre- and post-questionnaire surveys on students' expectations for oral communication classes are shown in Table 6.

Table 6: Survey Results of Students' Expectations for Oral English Classes

		Exactly Consistent	Consistent	Generally Consistent	Inconsistent	Definitely Inconsistent
Before teaching	EB1	78	16	4	4	0
	EB2	70	22	3	6	1
	EB3	78	13	5	6	0
After teaching	EA1	77	20	3	2	0
	EA2	79	13	9	1	0
	EA3	86	11	3	2	0

As shown in Table 6, prior to the action research, most students hoped for a more relaxed and lively oral classroom environment, with more teaching modalities and richer classroom activities. The actual effects of the systematic teaching approach were highly aligned with student expectations. Among students who desired a "relaxed and lively classroom," 100 students believed the system achieved this goal, while only 2 did not. Students expressed high satisfaction with the multimedia assistance provided by the oral communication system, with nearly all students acknowledging the multimedia teaching tools support (EA2) offered by the system. The level of activity richness was rated as "completely consistent" by 86 students, accounting for 84.31%, with dissatisfaction approaching 0.

IV. B. Analysis of test results

In order to further understand changes in students' oral English proficiency following the application of the oral English training system in teaching, this study conducted oral English tests before and after the experiment. The tests were scored on a 100-point scale, with (90, 100] indicating excellent, [80–90] indicating good, [60, 80) indicating passing, and (0, 60) indicating failing. The distribution of students' oral English test scores before and after the experiment is shown in Figure 7.

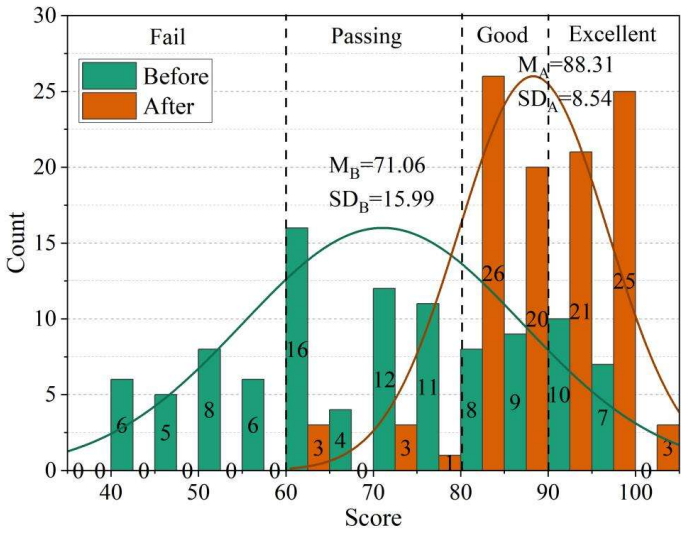

Figure 7: The distribution of students' English oral test scores

As shown in Figure 7, prior to the implementation of the English speaking system, the average speaking score of students was 71.06 ± 15.99 points. There were 25 students who failed, accounting for 24.51%, and only 17 students scored above 90 points, representing 16.67%. The majority of students were concentrated in the passing range of 60–80 points. After a semester of teaching using the speaking training system, students' scores improved to 88.31 ± 8.54, representing a significant increase of 24.25%. The number of failing students dropped to 0, with 49 students scoring above 90, nearly half of the total. Students scoring good and passing accounted for 45.10% and 6.86%, respectively. The significant improvement in oral scores indicates that the English oral training system designed in this paper has been effectively applied and is beneficial for enhancing students' oral performance.

V. Conclusion

This study designed and validated a deep learning-based English speaking training system. By integrating role information and historical influence vectors into a semantic understanding model, it significantly improved speaking comprehension in multi-turn dialogue scenarios. On the ATIS dataset, intent recognition accuracy reached 99.19%, surpassing the best baseline by 0.7%, and slot filling accuracy improved to 97.24%, leading by 1.1%. On the SNIPS dataset, the system achieved 98.38% intent accuracy and 98.19% slot filling accuracy, demonstrating the model's robustness across multiple domains.

The system performed exceptionally well in real teaching environments, with an average response latency of 1024.2ms, speech recognition accuracy of 98%, task completion rate of 98%, and pronunciation correction rate of 92%.

The average oral score of 102 students improved from 71.06 to 88.31 (+24.25%), the failure rate decreased from 24.51% to 0%, and the excellent rate (>90 points) increased from 16.67% to 48.04%; Learning attitudes improved significantly, with the number of students fully agreeing with the statement "speak English fluently" increasing by 95.1%, and the willingness to persist in daily training increasing by 252.4%.

Funding

This work was supported by Hebei Education Examinations Authority. It is a research outcome of 2024 Hebei Education Examinations and Enrollment Research Project: "Research on Digital Intelligence Empowering Foreign Language Oral Proficiency Tests in Higher Education Institutions" (Project No.: HBJK2024126).

References

[1] Lim, T. S. (2017). Verbal communication across cultures. Intercultural communication, 179-197.
[2] Shadiev, R., & Huang, Y. M. (2016). Facilitating cross-cultural understanding with learning activities supported by speech-to-text recognition and computer-aided translation. Computers & Education, 98, 130-141.
[3] Meifang, J., & Honghui, Z. (2020). Investigation and Analysis on the Oral Expression Competence of Students in Second-tier Universities--Based on China's Standards of English Language Ability. International Journal of Social Science, Innovation, & Educational Technologies, (14).
[4] Lifintsev, D., & Wellbrock, W. (2019). Cross-cultural communication in the digital age. Estudos em Comunicação, 1(28).
[5] Le, W., & Shuo, W. (2023). The cultivation of oral expression ability in the process of learning English in secondary schools. In SHS web of conferences (Vol. 179, p. 02015). EDP Sciences.
[6] Gabriel, M. L. H., Vergaray, J. M., Rivera-Arellano, E. G., & Flores, E. (2021). The Communicative Approach and Oral Expression in School: Theoretical Review. Turkish Journal of Computer and Mathematics Education, 12(6), 3241-3247.
[7] Styfanyshyn, I., & Kalymon, Y. (2020). Online practice for speaking English. Publishing house «UKRLOGOS Group», 124-132.
[8] Garcia-Ponce, E. E., Lengeling, M. M., Mora-Pablo, I., & Conaway Arroyo, L. M. (2023). Use of WhatsApp as a platform to promote English oral fluency and accuracy: A task repetition approach. Íkala, Revista de Lenguaje y Cultura, 28(1), 69-85.
[9] Al-Jarf, R. (2021). EFl speaking practice in distance learning during the coronavirus pandemic 2020-2021. International Journal of Research-GRANTHAALAYAH, 9(7), 179-196.
[10] Cruz-Ramos, M. D. L. M., & Herrera-Díaz, L. E. (2022). Assessment of Students' Oral Communicative Competence in English Through a Web Conferencing Platform. Profile Issues in TeachersProfessional Development, 24(1), 143-156.
[11] Mehdiyev, E. (2020). Using Role Playing in Oral Expression Skills Course: Views of Prospective EFL Teachers. International Online Journal of Education and Teaching, 7(4), 1389-1408.
[12] Altunkaya, H. (2018). The Impact of Activity-Based Oral Expression Course on Speech Self-Efficacy of Students. Journal of Education and Training Studies, 6(1), 137-150.
[13] Hu, W. (2024). A case study of incorporating digital technologies into audio-visual-oral English teaching in the context of digital humanities. In SHS Web of Conferences (Vol. 181, p. 04041). EDP Sciences.
[14] Jitpaisarnwattana, N. (2025). The effects of a personalised learning plan in a language MOOC on learners' oral presentation skills. The JALT CALL Journal, 21(2), 102601-102601.
[15] Dong, W., Pan, D., & Kim, S. (2024). Exploring the integration of IoT and Generative AI in English language education: Smart tools for personalized learning experiences. Journal of Computational Science, 82, 102397.
[16] Oh, E. Y., & Song, D. (2021). Developmental research on an interactive application for language speaking practice using speech recognition technology. Educational Technology Research and Development, 69(2), 861-884.
[17] Jiang, M. Y. C., Jong, M. S. Y., Lau, W. W. F., Chai, C. S., & Wu, N. (2021). Using automatic speech recognition technology to enhance EFL learners' oral language complexity in a flipped classroom. Australasian journal of educational technology, 37(2), 110-131.
[18] Paiva, V. L. M. D. O., & Ronaldo, C. G. J. (2019). Digital Tools for the Development of Oral Skills in English. Quarterly of Iranian Distance Education Journal, 2(1), 9-22.
[19] Chen, S. (2021). Design of internet of things online oral English teaching platform based on long-term and short-term memory network. International Journal of Continuing Engineering Education and Life Long Learning, 31(1), 104-118.
[20] Bauyrzhankyzy, A. A. (2025). THE ROLE OF ONLINE LEARNING PLATFORMS IN THE DEVELOPMENT OF SPOKEN LANGUAGE: NEW OPPORTUNITIES FOR LEARNING ENGLISH. Central Asian Scientific Journal, 58.
[21] Li, J., & Chen, H. (2021). Construction of case-based oral English mobile teaching platform based on mobile virtual technology. International Journal of Continuing Engineering Education and Life Long Learning, 31(1), 87-103.
[22] Kleftodimos, A., & Triantafillidou, A. (2023). The use of the video platform FlipGrid for practicing science oral communication. TechTrends, 67(2), 294-314.
[23] Chen, X. (2025). Interactive Strategies for English Oral Communication in a Network Environment. Education and Social Work, 1(1), 102-108.
[24] Xiuwen, Z., & Razali, A. B. (2021). An overview of the utilization of TikTok to improve oral English communication competence among EFL undergraduate students. Universal Journal of Educational Research, 9(7), 1439-1451.
[25] Xie, Y. (2023). Application of speech recognition technology based on machine learning for network oral English teaching system. International Journal of System Assurance Engineering and Management, 1-10.
[26] Liu, H. (2021). College oral English teaching reform driven by big data and deep neural network technology. Wireless Communications and Mobile Computing, 2021(1), 8389469.
[27] Chen, H., Tang, Y., & Du, J. (2024). Progress in using computer-assisted pronunciation training: The role of machine learning algorithms in the development of English oral skills. Journal of Computational Methods in Sciences and Engineering, 14727978251361030.
[28] Zhao, J., & Yang, J. (2025, March). English Oral Data Analysis and Feedback System Based on Support Vector Machine. In 2025 IEEE International Conference on Electronics, Energy Systems and Power Engineering (EESPE) (pp. 1092-1097). IEEE.
[29] Dubey, P., Dubey, P., Raja, R., & Kshatri, S. S. (2025). Bridging language gaps: The role of NLP and speech recognition in oral english instruction. MethodsX, 14, 103359.
[30] Zhan, X. (2025). Design and application of oral English training process combined with human-machine collaboration. Journal of Computational Methods in Sciences and Engineering, 14727978251361848.