Online-Journals.org

iJIM International Journal of Interactive Mobile Technologies

iJIM | eISSN: 1865-7923 | Vol. 19 No. 24 (2025) | OPEN ACCESS

https://doi.org/10.3991/ijim.v19i24.59477

PAPER

Application of Mobile Interactive Applications in College English Teaching from the Perspective of Intelligent Educational Technology

Ran Zhao (✉)

Hebei International Studies University, Shijiazhuang, China

zhaoranuk21@126.com

ABSTRACT

In the context of the challenges posed by personalization and effectiveness in college English teaching, the deep integration of intelligent educational technology and mobile learning has become a key solution. This paper aims to explore how advanced algorithms can be applied to mobile interactive applications to create an intelligent and immersive college English teaching environment. The study first designs an "Artificial intelligence (AI)-immersive task-based dialogue laboratory" as the core application scenario, where students engage in language learning by completing communication tasks in highly realistic virtual environments. To achieve natural interaction and precise feedback, this paper focuses on the development of two core technologies: First, a new continuous speech control method is proposed, optimizing matching efficiency and robustness through windowing limitations and weighting strategies, ensuring accurate understanding of students' natural spoken commands and smooth scene navigation. Second, a pronunciation quality evaluation system is constructed by integrating multiple speech features and adopting a multi-task learning framework, allowing for real-time, fine-grained diagnostics and feedback on pronunciation at the phoneme, word, and sentence levels. The study ultimately demonstrates that technology integration driven by educational scenarios can significantly enhance the interactive experience and teaching efficiency in college English learning, providing a practical solution for intelligent foreign language teaching.

KEYWORDS

intelligent educational technology, mobile interactive applications, college English teaching, continuous speech control, pronunciation quality evaluation, immersive learning

1 INTRODUCTION

With the rapid development of mobile internet [1, 2] and artificial intelligence (AI) technology [3–5], the education field is undergoing a profound intelligent transformation. In college English teaching [6, 7], the traditional "teacher-centered,

Zhao, R. (2025). Application of Mobile Interactive Applications in College English Teaching from the Perspective of Intelligent Educational Technology. *International Journal of Interactive Mobile Technologies (iJIM)*, 19(24), pp. 18–32. https://doi.org/10.3991/ijim.v19i24.59477

Article submitted 2025-09-03. Revision uploaded 2025-10-27. Final acceptance 2025-11-01.

textbook-centered, classroom-centered" model is increasingly showing its limitations and struggling to meet the students' needs for personalized, interactive, and practical learning. Mobile interactive applications [8, 9], with their ubiquity, convenience, and interactivity, offer a new path to solve this problem. However, most current educational applications are still at the initial stages of content digitalization [10] and exercise gamification [11], failing to fully tap into the potential of intelligent technologies, especially in the core area of oral communication skills training [12], and lacking advanced solutions that deeply integrate context and provide precise intelligent feedback. Against this backdrop, how to systematically integrate advanced intelligent educational technologies [13], especially continuous speech recognition and pronunciation quality evaluation algorithms, into mobile applications to build a highly immersive, instant feedback, and personalized college English learning environment has become an important issue that urgently needs to be explored and researched.

Although both academia and industry have long focused on the application of technology in foreign language teaching, existing research methods and solutions still have obvious flaws. For example, in the field of speech evaluation, most studies focus on isolated aspects, such as the system developed in [14], which only scores intonation, while the model in [15] mainly focuses on fluency. This "treating each problem individually" research paradigm leads to a limited evaluation dimension, making it difficult to provide comprehensive and holistic diagnostics of learners' oral performance. In speech control, many applications adopt static template matching [16] or simple keyword retrieval [17] methods, which appear rigid and fragile when processing natural, continuous dialogue and are unable to understand complex semantic intentions, severely limiting the naturalness and depth of human-computer interaction. In conclusion, existing technological solutions generally suffer from fragmented functionality, isolated scenes, and shallow feedback, failing to form an integrated whole of "intelligent dialogue interaction" and "multi-dimensional pronunciation diagnosis," making it difficult to support a complete and efficient immersive language learning environment.

To address the above deficiencies, this paper aims to research and construct an intelligent solution integrated into mobile applications, serving college English teaching. The main content of this research includes two core parts: First, application scenario description, namely designing and explaining an "AI-immersive task-based dialogue laboratory" teaching scenario, where students drive language learning by completing specific tasks in a highly realistic virtual context; second, key technology implementation, focusing on the two technical pillars supporting this scenario—an improved continuous speech control method to ensure natural and smooth dialogue interaction, and a pronunciation quality evaluation method to provide multi-layered, multi-dimensional feedback on phonemes, words, and sentences. The core value of this research lies in the fact that it is not a simple stacking of technologies but a deep practice of "scene-driven technology integration" guided by educational needs. By organically embedding advanced algorithms into carefully designed teaching contexts, this research provides a specific path that combines innovation, feasibility, and efficiency for the paradigm shift in college English teaching from "knowledge transmission" to "ability cultivation" and from "one-size-fits-all" to "personalized teaching."

2 APPLICATION SCENARIO DESCRIPTION

The "AI-immersive task-based dialogue laboratory" core application scenario proposed in this paper aims to address long-standing issues in college English

teaching, such as "mute English," lack of real-world context, and insufficient personalized feedback. This scenario simulates highly realistic cross-cultural communication contexts, such as check-in at an international airport, academic conference inquiries, or hotel check-in complaints. During the learning process, students are no longer engaged in isolated word repetition or sentence pattern practice; instead, as active participants, they must use continuous, natural spoken output to engage in complete, logical conversations with AI virtual characters to complete tasks. For example, in the "restaurant" scenario, students need to interact using natural sentences such as "I'm allergic to dairy, could you recommend a dish without cheese?" to handle everything from greeting to ordering, to dealing with sudden situations such as "food allergies." This design closely relies on the theories of situational learning and task-based teaching, anchoring language learning in concrete, meaningful tasks, thereby stimulating students' intrinsic motivation to learn and promoting the transformation of inert knowledge into transferable communicative skills.

To achieve the educational objectives mentioned above, this scenario deeply integrates continuous speech control and pronunciation quality assessment, linking them into a synergistic feedback organic whole. Continuous speech control technology serves as the engine driving the dialogue. It is based on an advanced end-to-end speech recognition model and natural language understanding module, which in real time analyzes the semantic intentions and contextual coherence of students' speech, ensuring that the AI character can provide appropriate responses, thus assessing students' comprehensive oral expression abilities. Meanwhile, the pronunciation quality assessment technology plays the role of an intelligent coach, utilizing deep neural network models to precisely analyze the phonemes, prosody, stress, and intonation of students' speech at the millisecond level. The deep integration of these two technologies is reflected in the system's ability to not only assess whether the conversation is "successful" but also diagnose its "quality." For example, if a student mispronounces "vegetarian" as /vɛdʒəˈtɛriən/, causing confusion for the AI character, the system will generate a comprehensive report after the conversation ends, clearly pointing out the pronunciation error and its potential impact on communication, and automatically push targeted minimal pair practice. This "intention-quality" dual-track evaluation mechanism perfectly reflects constructivist and mastery learning theories, providing students with multidimensional formative assessments beyond simple "correct/incorrect" judgments.

3 METHOD IMPLEMENTATION

3.1 Continuous speech control method

In the AI-immersive task-based dialogue laboratory constructed in this paper, the implementation of continuous speech control relies first on the precise "interpretation" of the raw speech signal. For this, continuous speech signal feature extraction is first performed, specifically extracting Mel-Frequency Cepstral Coefficient (MFCC) features. The human ear, as a nonlinear filter bank, is more sensitive to low-frequency sounds while being relatively less sensitive to high-frequency noise. MFCC replicates this mechanism perfectly by mapping the linear frequency spectrum to the Mel scale and applying a triangular bandpass filter bank. Specifically, in the complex acoustic environment of the dialogue laboratory, the continuous speech signal $a(v)$ entered by the student through a mobile device, after pre-emphasis compensation for high-frequency attenuation, is mapped

to the Mel frequency. Specifically, the frequency d of the continuous speech signal spectrum is first mapped to the Mel coordinate system:

$$D = 2595 \times \lg\left(1 + \frac{d}{700}\right) \tag{1}$$

The D is then input to the triangular bandpass filter, and the squared spectrum of the continuous speech signal is further calculated based on the filter output result $G_D(j)$:

$$\varphi_D(j) = \ln_2\left[\sum_{j=1}^{J}\left|A(l)\right|^2 G_D(j)\right] \tag{2}$$

Where J is the total number of triangular bandpass filters. Finally, the MFCC of the continuous speech signal is obtained based on the following formula:

$$b(v) = \sum_{j=1}^{J}\varphi_D(j)COS\left[v(j-0.5)\frac{\tau}{J}\right] \tag{3}$$

Where M represents the order of the MFCC. The above operations are not performed with average effort but are strategically focused on the low-frequency region from 500Hz to 2000Hz, which carries the semantic key, while naturally suppressing high-frequency noise interference. This enables the subsequent speech recognition engine to obtain a "purified" and "highlighted" feature representation, allowing for more accurate understanding of the speech content and intention when students perform tasks such as "check-in" or "ordering food."

College English learners' speech generally exhibits "non-standard" characteristics, such as inaccurate pronunciation and unstable rhythm, which can be seen as noise and outliers from an algorithmic perspective. Traditional dynamic time warping algorithms treat all feature points equally, making it easy to misalign the entire sentence sequence due to severe distortion of individual phonemes, which leads to misjudging the semantic intention, such as confusing the mispronounced "ship" with "sheep," thus disrupting the dialogue logic. To improve the system's fault tolerance and robustness in educational scenarios, this paper adopts a weighted dynamic time warping strategy for improvement. The core principle of this strategy is to introduce a weight factor when calculating the distance between sequences, assigning different levels of importance to different frames or dimensions in the MFCC feature sequence. Specifically, the system can allocate weights based on the following strategy: 1) Assigning higher weight to vowel segments and stable sound segments that carry key semantic information to ensure the alignment accuracy of the sentence backbone; 2) Assigning lower weight to anomalous feature dimensions with large variance, introduced by the individual learner's pronunciation habits, in order to reduce their interference. Specifically, assuming the distance between the g-th feature point in the improved $b_u(v)$ and the h-th feature point in the $b_k(l)$ is represented by $f_\mu(g, h)$, the weight factor is represented by $\mu(g, h)$, and the distance between the g-th feature point in the unmodified $b_u(v)$ and the h-th feature point in the $b_k(l)$ is represented by $f[b_{u,g}(v), b_{k,h}(l)]$, the distance function expression of the improved dynamic time warping algorithm is:

$$f_\mu(g, h) = \mu(g, h) \cdot f[b_{u,g}(v), b_{k,h}(l)] \tag{4}$$

Through this weighting mechanism, the system can "intelligently" ignore unimportant and variable details in the learner's pronunciation while firmly grasping

the key parts that determine semantic understanding, thereby accurately understanding the student's instructions in the task even when the pronunciation is not perfect, achieving the transformation from a "strict pronunciation evaluator" to a "sympathetic dialogue partner." This is the essence of educational technology enabling a personalized, supportive learning environment.

In the AI-immersive task-based dialogue laboratory, the ultimate goal of continuous speech control is to accurately understand the student's oral instructions and map them to specific teaching interaction behaviors. The core of this process is to match the student's speech sequence $b(v)$, which has been optimized through MFCC feature extraction, with a series of preset speech instruction templates $b^k(l)$ that represent different communicative functions. These templates are not simple mechanical commands like "go forward," "stop," but encapsulate typical expressions with specific pragmatic intentions, such as "making a request," "apologizing," "inquiring for information," or "arguing," among other diversified templates. To achieve efficient and accurate matching, the system first divides the long speech sequence $b(v)$ into T sub-sequences based on semantic boundaries, then applies the improved dynamic time warping algorithm to calculate the weighted accumulated distance F_k between each pair of sub-sequences $b_t(v)$ and $b_t^k(l)$. Assuming that the Euclidean distance between $b_{t,g}(v)$ and $b_{t,h}^k(l)$ is represented by $f[b_{t,g}(v), B_{t,h}^k(l)]$, the accumulated distance expression is:

$$F_k = \sum_{|o-w|\le e} \sum_{t=1}^{T} \sum_{g=1}^{o} \sum_{h=1}^{w} \mu(g,h) \cdot f[b_{t,g}(v), B_{t,h}^k(l)] \tag{5}$$

Further through the formulation of recognition rules based on the minimum accumulated distance, the system can determine the best matching template from K candidate templates, thereby completing the transformation from the student's continuous, variable, and possibly non-standard pronunciation acoustic signal to highly structured teaching semantic instructions. Specifically, let the recognition threshold be represented by $\hat{F}$, the specific expression of the recognition rule is as follows:

$$\begin{cases} F_k \ge \hat{F}, \textit{The current speech is the } k-th \textit{ speech command} \\ F_k < \hat{F}, \textit{The current speech is not the } k-th \textit{ speech command} \end{cases} \tag{6}$$

The successful recognition of continuous speech instructions directly triggers the intelligent control feedback loop in the laboratory scene's teaching process. The "control" here refers to the advancement of the virtual dialogue context, the generation of teaching feedback, and the guidance of the learning path. Once the student's speech command is recognized, the system immediately transmits this instruction to the two core "executive agencies," the scene engine and the dialogue manager. The scene engine is responsible for updating the virtual environment's state based on the instruction, and more importantly, the dialogue manager generates multi-dimensional formative teaching feedback based on the results of this speech recognition. If the student uses a very authentic expression with clear pronunciation, the system will give positive affirmation through the virtual character. If the matching confidence is low or the recognized intention does not match the context, the system will trigger a "clarification request" from the virtual character, guiding the student to express themselves in a different way. This adaptive interaction based on recognition results constitutes a powerful teaching control loop, which not only executes the commands but also evaluates and promotes the student's language use ability through the process of executing the commands.

3.2 Pronunciation quality evaluation method

In the AI-immersive task-based dialogue laboratory constructed in this paper, the core goal of pronunciation quality evaluation goes beyond traditional single-dimensional scoring. It aims to conduct a multi-dimensional, in-depth "teaching diagnosis" of students' oral output. Traditional models focus only on aspects like intonation and fluency, akin to checking only a single health indicator without giving a comprehensive health report. This paper chooses to introduce the glottalization of phoneme and tone (GOPT) model, i.e., adopting a multi-task learning framework to jointly model and evaluate the three levels of phonemes, vocabulary, and sentences in speech.

Specifically, the model uses glottalization of phoneme (GOP) features to accurately locate the pronunciation accuracy of each phoneme, which is the micro foundation for evaluation. At the word level, it can assess whether the stress pattern of multi-syllable words is correct. At the sentence level, it performs a comprehensive analysis of the overall prosody, rhythm, and intonation. The GOP features are obtained from the log of phoneme posterior ratios and the log posterior ratio. Assuming the standard phoneme is represented by o, the input observation is represented by p, and the start and end frame indices are represented by s_t and s_r, the log phoneme posterior ratio calculation formula is as follows:

$$e_{DSYS}(o) = \ln o(o \mid p; s_t, s_r) \approx \frac{1}{s_r - s_t + 1} \sum_{s=s_t}^{s_r} \ln(o \mid p_s) \tag{7}$$

The log posterior ratio calculation formula is as follows:

$$e_{DS}(o_k, o_u) = \ln o(o_k \mid p; s_t, s_r) - \ln o(o_u \mid p; s_t, s_r) \tag{8}$$

The GOP feature expression is as follows:

$$[e_{DSYS}(o_1) \ldots e_{DSYS}(o_M), r_{DS}(o_1 \mid o_u) \ldots e_{DS}(o_M \mid o_u)]^T \tag{9}$$

The innovation of this design lies in that it completes a comprehensive analysis from "local correctness" to "global expressiveness" through a unified Transformer model. For example, when a student says, "I'm very interested in this opportunity" in the dialogue, the GOPT model not only diagnoses the phoneme-level issue of the /v/ sound in "very," but also assesses whether the stress of "opportunity" falls on the second syllable rather than the first, and finally evaluates whether the overall intonation when expressing "interest" is enthusiastic and natural at the sentence level.

College English learners come from different dialect areas, and their pronunciation errors vary greatly, changing continuously throughout the learning process. This requires the evaluation model to possess strong generalization ability and a certain level of interpretability to provide genuinely valuable feedback for teaching. To this end, this paper introduces the Squeezeformer model. Squeezeformer can dynamically adjust the constraint strength during the training process, effectively preventing overfitting, enabling the model not only to perform well on standard, clear speech data but also to accurately evaluate "real learner speech" with various accents, interruptions, or non-standard pronunciation, thus ensuring stability and reliability in frontline teaching. At the same time, Squeezeformer retains and optimizes the self-attention mechanism, which allows the model to clearly "inform" the system of the specific phonemes or word fragments it focused on when evaluating a sentence's pronunciation. For example, when the system points out that the student's pronunciation of the word "opportunity" is poor, teachers or students themselves can analyze the feedback report by visualizing the attention weights and

discovering that the model mainly made the judgment based on the unclear pronunciation of the stressed syllable "tu" in that word. This interpretability of the prediction results transforms pronunciation evaluation from an abstract score-giving "black box" to a "transparent coach" that can pinpoint the specific issues, greatly enhancing the teaching value and credibility of the feedback, allowing students to know the reason behind the issue and make targeted improvements.

In the AI-immersive task-based dialogue laboratory constructed in this paper, the core of pronunciation quality evaluation is to build a multi-dimensional perceptual system capable of comprehensively and accurately diagnosing students' oral problems. Its basic principle begins with the deep fusion of multi-source heterogeneous features. The system does not rely on a single type of information but instead cooperates to utilize the advantages of three features: the GOP features based on the acoustic model accurately depict the phoneme-level pronunciation accuracy, providing a solid acoustic foundation for evaluation while the deep features extracted from self-supervised pre-trained models such as HuBERT and WavLM contain rich prosody, rhythm, and broader acoustic context information. By concatenating and fusing these three features with specific weights, the system obtains an 86-dimensional enhanced feature vector that can both anchor micro-phoneme errors and perceive macro-prosody characteristics. Subsequently, this feature is input into the improved Squeezeformer-MR encoder along with the standard phoneme embedding and position encoding. This model, through the combination of its self-attention mechanism and convolutional structure, can simultaneously capture long-range global dependencies and short-range local patterns in the speech sequence, thus enabling hierarchical joint modeling from phonemes to words, to sentences. Figure 1 shows the structure of the Squeezeformer-MR. The specific process expression of this structure is as follows:

$$R = (R_{GOP}, uR_{HuBERT}, kE_{WavLM}) \tag{10}$$

$$R' = BN(Dense(R)) \tag{11}$$

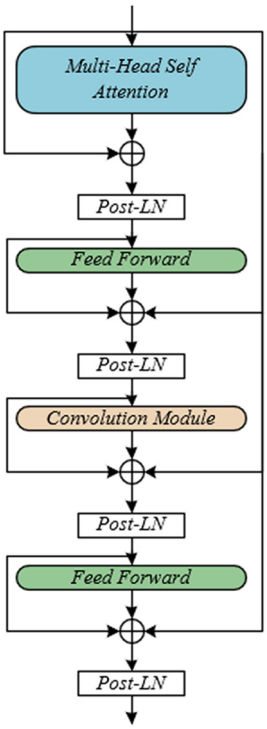

Fig. 1. Squeezeformer-MR structure

The ultimate goal of this evaluation system is to transform its powerful analytical capabilities into a real driving force for promoting the teaching process. Its principle reflects a teaching adaptability design integrating multi-task learning and efficient architecture. In terms of output, the model, through the five pre-trained [cls] tokens, concurrently performs multiple evaluation tasks: these tokens, after being processed by the Squeezeformer-MR encoder, output to different regression heads corresponding to five key teaching indicators, such as fluency, prosody, stress, completeness, and overall intelligibility. This multi-task architecture means that the system can complete a comprehensive "check-up" of the learner's pronunciation in one forward propagation, rather than calculating each indicator sequentially, which provides algorithmic support for achieving real-time feedback with low latency on mobile devices. Squeezeformer-MR itself, by reducing the number of parameters and introducing residual links and normalization, ensures the model's efficient and stable operation on mobile devices, meeting the stringent real-time requirements of immersive dialogues.

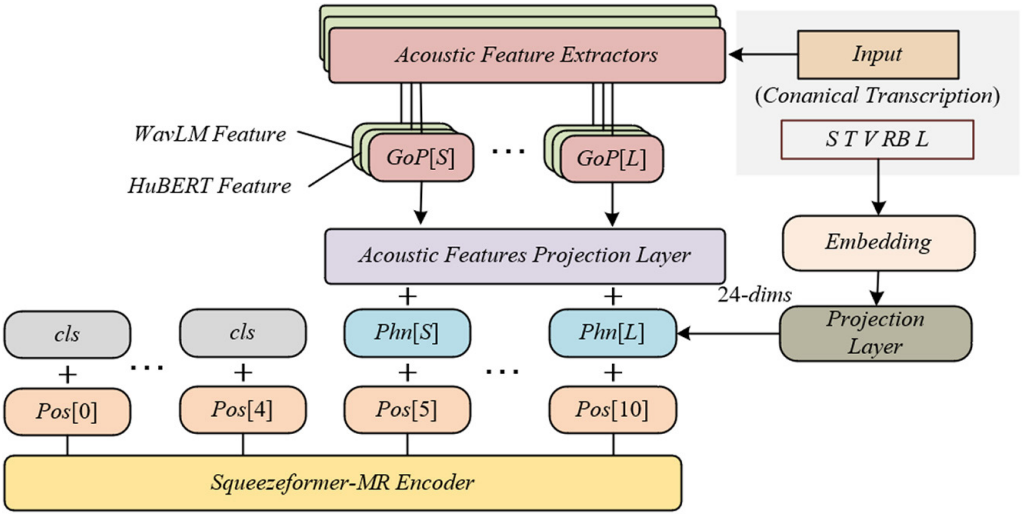

Fig. 2. Pronunciation quality evaluation model framework

Finally, after the dialogue task is completed, the system can immediately generate a structured diagnostic report, not only providing a score but also clearly pointing out the direct causes of the diagnostic results and directly linking to targeted practice modules. This fully reflects, from the perspective of educational technology, the seamless integration of assessment, diagnosis, and teaching intervention through advanced algorithms, ultimately realizing the core purpose of personalized, mastery-based learning. Figure 2 presents the overall framework of the pronunciation quality evaluation model.

4 EXPERIMENTAL RESULTS AND ANALYSIS

To validate the reliability of continuous speech control as the core system driver, we compared the recognition performance of different models. As shown in Table 1, in the ideal quiet laboratory environment, the end-to-end Transformer model performs best in terms of word error rate, but its sentence error rate is still

higher than the improved dynamic time warping (DTW) method proposed in this paper, and its average response time far exceeds that of the proposed method, making it difficult to meet the low-latency requirements for real-time dialogue. In a noisy classroom environment, which is closer to real teaching scenarios, the advantage of the proposed method becomes more evident: its word error rate and sentence error rate are significantly lower than the traditional DTW algorithm, while compared to the computationally intensive Transformer model, the word error rate (WER) and sentence error rate (SER) are reduced by 22.4% and 23.3%, respectively, and the response time is controlled at a very low level of 305 ms. This result fully demonstrates that the improved DTW algorithm proposed in this paper, by achieving an excellent balance between computational efficiency and noise robustness, provides a continuous speech recognition solution that is responsive and highly resistant to interference, ensuring the smoothness and naturalness of human-computer teaching dialogue in the "AI-Immersive Task-based Dialogue Laboratory."

Table 1. Comparison of continuous speech recognition accuracy

Model/Method	Test Environment	Word Error Rate (WER)	Sentence Error Rate (SER)	Average Response Time (ms)
Traditional DTW Algorithm	Quiet Laboratory	18.5%	35.2%	320
	Noisy Classroom	35.8%	62.1%	350
End-to-End Transformer Model	Quiet Laboratory	12.1%	28.5%	580
	Noisy Classroom	25.4%	50.3%	610
Proposed Method	Quiet Laboratory	14.3%	25.8%	280
	Noisy Classroom	19.7%	38.6%	305

To confirm that the pronunciation evaluation system proposed in this paper provides more accurate and comprehensive feedback compared to traditional methods, we performed a multi-dimensional comparison with mainstream baseline models. The data in Table 2 clearly show that the proposed GOPT model leads in almost all core metrics. Its Pearson correlation with expert ratings reaches 0.89, far higher than the baseline model, and its RMSE is as low as 0.98, which proves that the GOPT model's scoring has a high consistency with human expert judgment, making it highly reliable. More importantly, in specific diagnostic tasks, the GOPT model significantly outperforms in terms of phoneme, word-level stress, and sentence-level fluency recognition accuracy, especially in sentence-level fluency, where it improved by nearly 20 percentage points compared to the powerful DNN baseline. This strongly demonstrates the effectiveness of multi-feature fusion and multi-task joint modeling, enabling the model to collaboratively understand all information from micro-phonemes to macro-prosody. Although the inference time of GOPT is slightly higher than that of the simplest model, 42ms is fully sufficient to meet the real-time feedback requirement. In conclusion, this experimental result strongly supports the core value of this research: the GOPT model can serve as a "real-time coach," providing learners with pronunciation quality evaluation feedback that is highly accurate, multi-dimensional, and practically useful in teaching.

Table 2. Comparison of pronunciation quality evaluation model performance

Evaluation Model	Pearson Correlation with Expert Rating	Prediction Error	Phoneme-Level Error Recognition Accuracy	Word-Level Stress Error Recognition Accuracy	Sentence-Level Fluency Rating Accuracy	Model Inference Time (ms)
Baseline Model: GOP + Logistic Regression	0.65	1.85	70.2%	58.5%	55.1%	<10
GOP + DNN	0.72	1.52	75.8%	65.3%	63.7%	35
Proposed Model: GOPT (Squeezeformer-MR)	0.89	0.98	84.5%	78.9%	82.4%	42

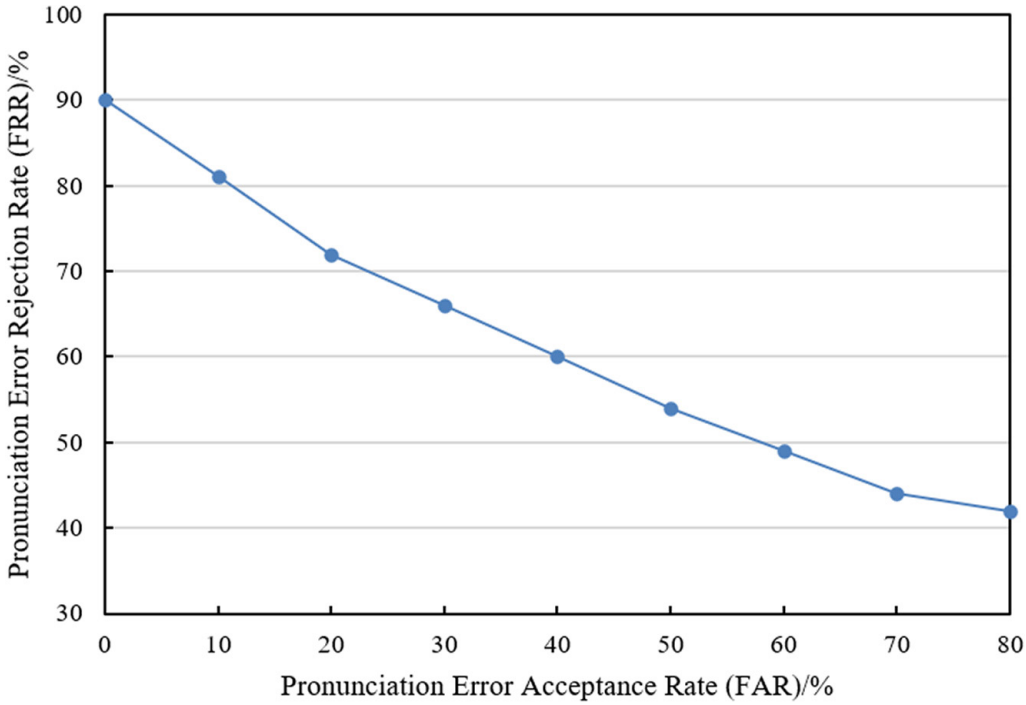

Fig. 3. Performance evaluation of pronunciation error acceptance rate and rejection rate

To verify the reliability and applicability of the pronunciation quality evaluation system built in this paper as a "real-time coach," we conducted a rigorous performance evaluation of its core discrimination model. The ROC curve data in Figure 3 reveal the system's precise trade-off ability when discriminating between "correct pronunciation" and "incorrect pronunciation." Specifically, when the pronunciation error acceptance rate rises from 0% to 80%, the pronunciation error rejection rate steadily drops from 90% to 42%. This set of data clearly illustrates the adjustable range of system performance: if the system is set to be extremely strict, it can ensure that all errors are captured, but it will misjudge many correct pronunciations, which may overly undermine the student's confidence; on the other hand, if the system is set too leniently, it may maximize student encouragement but miss most errors, rendering feedback meaningless. For teaching applications, the key is to select a balanced point. For instance, with a false acceptance rate (FAR) of around 30%, the system can effectively capture most pronunciation errors without overly suppressing the student's desire to express themselves due to excessive strictness. This adjustable, high-performance discrimination characteristic fundamentally

ensures the scientific and pedagogical soundness of the feedback mechanism in the "AI-Immersive Task-based Dialogue Laboratory," allowing it to flexibly adjust feedback strategies based on different teaching stages or student levels, truly becoming an intelligent, supportive "coach."

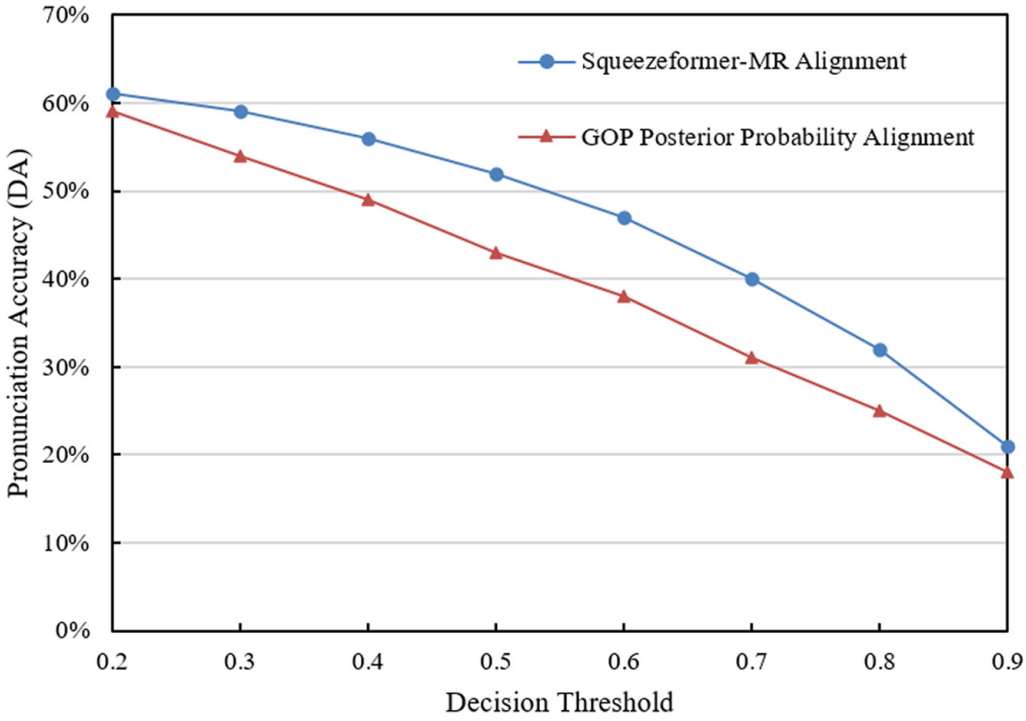

Fig. 4. Comparison of pronunciation accuracy (DA) under different alignment methods

To demonstrate the advancement of the Squeezeformer-MR model used in this paper for pronunciation evaluation tasks, we compared it with the traditional GOP posterior probability alignment method. The data in Figure 4 strongly confirm that, under different decision thresholds, the proposed Squeezeformer-MR alignment method consistently outperforms the traditional baseline method in terms of pronunciation accuracy. The experimental data show that at all tested threshold points, the DA value of Squeezeformer-MR is higher than that of the GOP posterior probability alignment, especially at a threshold of 0.4, where the former's decision accuracy (DA) is 56%, while the latter's is only 49%, showing a significant advantage of seven percentage points. This advantage is not coincidental; it arises from the powerful contextual modeling ability of the Squeezeformer-MR model, which, when judging a single phoneme pronunciation, can fully utilize word- and sentence-level contextual information, thus making more precise and human-like judgments than the GOP method, which only relies on local acoustic posterior probabilities. More importantly, the DA curve of Squeezeformer-MR drops more gently, indicating that its performance is more robust to changes in decision thresholds. This stability is crucial for teaching applications aimed at a large number of non-standard pronunciation learners, meaning that the system can provide consistent, reliable evaluation services without the need for complex parameter tuning for different accents, significantly enhancing the system's universality and practical value in large-scale university English teaching. This is the core competitive advantage of the technological solution proposed in this study.

Table 3. Analysis of learner pronunciation progress trajectory

Learner Group	Evaluation Dimension	Pre-Test (Week 1)	Mid-Test (Week 4)	Post-Test (Week 8)	Progress (Percentage Points)
High-level Group ($n = 10$)	Overall Pronunciation Score	82.5	85.1	87.2	+4.7
	Vowel Accuracy	90.3	91.5	92.8	+2.5
	Consonant Accuracy	88.1	89.0	89.9	+1.8
	Word-level Stress Accuracy	80.5	83.2	86.0	+5.5
	Sentence-level Fluency	78.2	82.5	85.1	+6.9
	Average Weekly Practice Frequency	4.2	3.8	3.5	–0.7
Mid-level Group ($n = 15$)	Overall Pronunciation Score	65.3	72.8	78.9	+13.6
	Vowel Accuracy	72.5	78.1	82.0	+9.5
	Consonant Accuracy	70.1	75.3	80.5	+10.4
	Word-level Stress Accuracy	58.9	68.5	76.2	+17.3
	Sentence-level Fluency	55.7	66.0	74.8	+19.1
	Average Weekly Practice Frequency	3.5	4.1	4.3	+0.8
Low-level Group ($n = 15$)	Overall Pronunciation Score	48.7	60.1	70.5	+21.8
	Vowel Accuracy	55.2	65.8	74.3	+19.1
	Consonant Accuracy	52.1	62.5	72.1	+20.0
	Word-level Stress Accuracy	40.5	53.2	65.7	+25.2
	Sentence-level Fluency	38.9	52.0	64.9	+26.0
	Average Weekly Practice Frequency	2.8	3.9	4.5	+1.7

To explore the long-term teaching effectiveness of the system for learners of different proficiency levels, we tracked their pronunciation progress. The data in Table 3 clearly reveal a key conclusion: learners at all levels made significant progress, with the lowest-level group showing the greatest improvement. The overall pronunciation score of the low-level group increased by 21.8 percentage points, far surpassing both the mid-level and high-level groups. In specific dimensions, the most notable progress across all groups was observed in word-level stress and sentence-level fluency, which highlights the system's emphasis on suprasegmental features and feedback in immersive task-based dialogue. Furthermore, the positive correlation between behavioral data and progress is compelling: the low-level group's motivation was greatly stimulated after receiving positive feedback, leading to an increase in the average weekly practice frequency from 2.8 to 4.5 times, whereas the high-level group, due to a stronger foundation, naturally experienced a decrease in practice frequency. This strongly indicates that the system,

by providing precise and positive feedback, is particularly effective in stimulating the learning motivation of students with weaker foundations and helping them make breakthroughs in core challenges of foreign language learning. This result robustly supports the core value of this research, namely that the mobile interactive application developed can achieve truly personalized, mastery-based learning and serves as an efficient and equitable intelligent solution for university English teaching.

5 CONCLUSION

This research systematically explored the innovative application of intelligent educational technology in university English teaching, with the core contribution being the development and validation of a mobile interactive solution that integrates the "AI-Immersive Task-based Dialogue Laboratory" scenario, improved continuous speech control, and multi-dimensional pronunciation quality assessment. The research results indicate that, through the improved DTW algorithm, the system achieves continuous speech recognition with a WER below 20% and an SER below 39% in real classroom noise environments, ensuring the natural and smooth flow of human-computer teaching dialogue. Meanwhile, the GOPT model based on multi-feature fusion and the Squeezeformer-MR architecture achieves a high Pearson correlation of 0.89 with expert ratings in pronunciation evaluation tasks and provides accurate diagnostics at the phoneme, word-level stress, and sentence-level fluency. The teaching experiment data further confirm that this application significantly improves students' speaking abilities, particularly for mid- and low-level learners, with overall pronunciation score improvements ranging from 13.6% to 21.8%, and effectively stimulates students' intrinsic learning motivation. The main value of this research lies in its successful transformation of cutting-edge algorithms into practical teaching productivity, demonstrating that the "scenario-driven technology integration" paradigm can build an intelligent learning environment that combines technological robustness, teaching effectiveness, and user-friendly experience, offering a clear and feasible path to address the challenges of university English speaking instruction.

However, this study still has certain limitations. First, the system's performance optimization and experiments were mainly conducted in predefined specific task scenarios, and there are still shortcomings in supporting and assessing open and creative spoken expressions. Second, while the research sample is somewhat representative, there is still room for expansion in both scale and diversity to further validate the system's universality. Furthermore, the current system primarily focuses on students' "learning" and has not fully explored how to transform the rich learning data into insights to assist teachers in classroom management and teaching decision-making. Based on this, future research directions should include: first, expanding the openness and complexity of tasks, introducing higher-level dialogue scenarios such as debates and speeches, and developing corresponding evaluation algorithms; second, deepening the personalized adaptive mechanism so that the system can dynamically adjust task difficulty and feedback strategies based on learners' real-time performance, achieving truly "differentiated instruction"; third, constructing a teacher dashboard to visualize student data and empower teachers to make precise instructional interventions. Through continuous iteration, this research paradigm is expected to become an indispensable core component of future intelligent foreign language teaching ecosystems.

6 REFERENCES

[1] A. D. Samala, D. Mhlanga, L. Bojic, N.-J. Howard, and D. P. Coelho, "Blockchain technology in education: Opportunities, challenges, and beyond," *International Journal of Interactive Mobile Technologies (iJIM)*, vol. 18, no. 1, pp. 20–42, 2024. https://doi.org/10.3991/ijim.v18i01.46307

[2] S. Sakulwichitsintu, "Mobile technology – an innovative instructional design model in distance education," *International Journal of Interactive Mobile Technologies (iJIM)*, vol. 17, no. 7, pp. 4–31, 2023. https://doi.org/10.3991/ijim.v17i07.36457

[3] E. Kyriaki, E. Giama, and I. Theodoridou, "Machine learning, artificial intelligence and digital twins: An up-to-date review analysis of the latest-era technologies in the urban building sector," *International Journal of Sustainable Energy*, vol. 44, no. 1, p. 2544238, 2025. https://doi.org/10.1080/14786451.2025.2544238

[4] A. Tovmasyan, N. Weinstein, and B. Mittelstadt, "Values tensions and values tradeoffs in the development of healthcare artificial intelligence technology: A conceptual model of decisions to create trustworthy technology," *Social Influence*, vol. 20, no. 1, p. 2478940, 2025. https://doi.org/10.1080/15534510.2025.2478940

[5] A. Shahini *et al.*, "A systematic review for artificial intelligence-driven assistive technologies to support children with neurodevelopmental disorders," *Information Fusion*, vol. 124, p. 103441, 2025. https://doi.org/10.1016/j.inffus.2025.103441

[6] J. Tang, "College English classroom teaching method and practice based on big data technology," *Journal of Computational Methods in Sciences and Engineering*, 2025. https://doi.org/10.1177/14727978251366529

[7] L. Cheng, "Big data analysis in college English teaching mobile information system based on advanced computational discourse analysis," *Journal of Computational Methods in Sciences and Engineering*, 2025. https://doi.org/10.1177/14727978251366526

[8] F. Tian, X. Zhang, J. Liang, and Z. Yang, "Bidirectional service function chain embedding for interactive applications in mobile edge networks," *IEEE Transactions on Mobile Computing*, vol. 23, no. 5, pp. 3964–3980, 2023. https://doi.org/10.1109/TMC.2023.3282645

[9] R. Dwivedi, A. D. Goel, V. Vyas, P. P. Sharma, P. Singh, and K. Singh, "Quality assessment of interactive mobile application–maternal and infant care for tribal birth attendants," *Indian Journal of Public Health*, vol. 68, no. 2, pp. 305–309, 2024. https://doi.org/10.4103/ijph.ijph_740_23

[10] Z. K. Cenk and S. Arslan Selçuk, "Scientific mapping of digitalization in architectural education for sustainability," *VITRUVIO-International Journal of Architectural Technology and Sustainability*, vol. 10, no. 1, p. 22911, 2025. https://doi.org/10.4995/vitruvio-ijats.2025.22911

[11] N. Sribundit *et al.*, "Innovation in teaching and learning: Gamification toward enhancing the performance of entrepreneurial skills and leadership skill in pharmacy student," *Currents in Pharmacy Teaching and Learning*, vol. 17, no. 5, p. 102301, 2025. https://doi.org/10.1016/j.cptl.2025.102301

[12] M. E. Lunz and P. G. Bashook, "Relationship between candidate communication ability and oral certification examination scores," *Medical Education*, vol. 42, no. 12, pp. 1227–1233, 2008. https://doi.org/10.1111/j.1365-2923.2008.03231.x

[13] A. Khamis, "Smart mobility education and capacity building for sustainable development: A review and case study," *Sustainability*, vol. 17, no. 17, p. 7999, 2025. https://doi.org/10.3390/su17177999

[14] F. López-Calatayud and J. Tejada, "Self-regulation strategies and behaviors in the initial learning of the viola and violin with the support of software for real-time instrumental intonation assessment," *Research Studies in Music Education*, vol. 46, no. 1, pp. 48–65, 2024. https://doi.org/10.1177/1321103X221128733

[15] S. Detey, L. Fontan, M. Le Coz, and S. Jmel, "Computer-assisted assessment of phonetic fluency in a second language: A longitudinal study of Japanese learners of French," *Speech Communication*, vol. 125, pp. 69–79, 2020. https://doi.org/10.1016/j.specom.2020.10.001

[16] H. M. Smith, K. L. Ritchie, T. S. Baguley, and N. Lavan, "Face and voice identity matching accuracy is not improved by multimodal identity information," *British Journal of Psychology*, vol. 116, no. 2, pp. 367–385, 2025. https://doi.org/10.1111/bjop.12757

[17] M. Biran, G. Ben-Or, and H. Yihye-Shmuel, "Word retrieval in aphasia: From naming tests to connected speech and the impact on well-being," *Aphasiology*, vol. 38, no. 4, pp. 738–757, 2024. https://doi.org/10.1080/02687038.2023.2228017

7 AUTHOR

Ran Zhao graduated from The University of Nottingham in 2021 is working at School of Teacher Education, Hebei International Studies University, Shijiazhuang 050000, China, and is engaged in the Motivation of the Second Language and the Directed Motivational Currents (DMCs) (E-mail: zhaoranuk21@126.com).