Building an interactive learning space driven by generative artificial intelligence: personalized English learning experience

Xinyu Ni[1,a,*]

[1]*University of Central Missouri, 116 W South St, Warrensburg, MO 64093, United States(US)*

[a]*qinghui.xu.sg@gmail.com,,*

[*]*corresponding author*

ABSTRACT:*English language education is undergoing a transformative shift, propelled by advancements in technology. This research explores the integration of Generative Artificial Intelligence (Generative AI) and interactive learning environments in the context of English language education, with a focus on developing a personalized oral assessment method. The proposed method leverages Generative AI's language generation capabilities within an interactive learning space to create a dynamic, adaptive learning environment. The study addresses historical challenges in traditional teaching methodologies, emphasizing the need for AI-driven approaches. The research objectives encompass a comprehensive exploration of the historical context, challenges, and existing technological interventions in English language education. A novel, technology-driven oral assessment method is designed, implemented, and rigorously evaluated using datasets such as Librispeech and L2Arctic. The ablation study investigates the impact of training dataset proportions and model learning rates on the method's performance. Results from the study highlight the importance of maintaining a balance in dataset proportions, selecting an optimal learning rate, and considering model depth to achieve optimal performance.*

Key words:*Generative AI;Interactive Learning;English Language Education;Personalized Learning;Oral Assessment*

1. Introduction

English language education is undergoing a significant transformation, driven by the need to adapt traditional teaching methodologies to the diverse and evolving needs of learners. As the global education landscape continues to evolve(Baidoo-Anu, D. & Ansah, L. O. 2023), there is growing recognition of the need to personalize pedagogical approaches to better address these challenges. This research explores the integration of cutting-edge technologies, specifically Generative Artificial Intelligence (Generative AI), within interactive learning environments to create dynamic, adaptive spaces for English language education(Graves, A., Fernandez, S. & Gómez, F. 2006). By harnessing the capabilities of AI to foster real-time interactions(Wang, Y., Skerry-Ryan, R. J., Stanton, D., et al. 2017) and personalized feedback(Liu, R., Chen, X. & Wen, X. 2020), this approach aims to provide learners with more tailored, engaging, and effective language learning experiences(Zhao, G., Sonsaat, S., Silpachai, A., et al. 2018).

English language education has faced the challenge of aligning teaching methods with the individualized nature of language acquisition(Lee, A., Chen, N. F. & Glass, J. 2016). Traditional approaches, often based on standardized curricula and assessments, have struggled to accommodate the wide variety of linguistic proficiency levels, cultural backgrounds, and learning preferences among students(Castro, G. P. B., Chiappe, A., Rodríguez, D. F. B.

& Sepulveda, F. G. 2024). This lack of personalization has led to disparities in educational outcomes and highlighted the need for innovative pedagogical solutions. As the educational landscape becomes increasingly globalized and diverse(Wang, Y., Skerry-Ryan, R. J., Stanton, D., et al. 2017), the limitations of conventional methods are becoming more apparent, underscoring the need for a shift toward more flexible, student-centered learning environments(Moritz, N., Hori, T. & Le, J. 2020).

In response to these challenges, interactive learning spaces, powered by Generative AI, offer an exciting opportunity to revolutionize English language education. By embedding AI within these dynamic learning environments, students can engage with content in ways that adapt to their unique linguistic, cognitive, and cultural needs(Filippetti, S., Sbardella, T. & Montanucci, G. 2024). Generative AI, with its ability to create contextually relevant and personalized content, can simulate interactive dialogues, provide real-time feedback(Panayotov, V., Chen, G. & Povey, D. 2015), and dynamically adjust to each learner ' s pace and proficiency. This contrasts with traditional AI, which typically processes data to map input to output, by instead generating new, customized learning experiences that evolve alongside the learner(Chimbga, B. 2023).

The integration of Generative AI within interactive learning environments offers the potential not only to improve language proficiency but also to enhance adaptability, inclusivity, and cultural sensitivity(Ayeni, O. O., Al Hamad, N. M., Chisom, O. N., Osawaru, B. & Adewusi, O. E. 2024). In this research, we aim to explore how this technological convergence can address the limitations of traditional teaching methods and create more personalized, engaging, and effective English language education experiences(Wei, L. 2023). Through this approach, the goal is to foster a more responsive and learner-centered educational ecosystem that supports the diverse needs of English language learners across the globe(Higuchi, Y., Watanabe, S., Chen, N., et al. 2020).

Generative Artificial Intelligence (AI), including technologies like Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs), has the ability to learn the underlying structures of data and generate new, realistic samples(Yekollu, R. K., Bhimraj Ghuge, T., Sunil Biradar, S., Haldikar, S. V. & Farook Mohideen Abdul Kader, O. 2024). This capability has led to innovative applications across fields such as art, music, and storytelling, but also raises ethical concerns regarding the misuse of generated content, such as deepfakes, and its broader impact on society and culture(Yan, B. C., Wu, M. C., Hung, H. T., et al. 2020).

In the context of English language education, the integration of Generative AI within interactive learning spaces holds immense promise. These spaces, designed to engage learners actively, provide dynamic environments where AI can simulate real-time interactions, personalize content, and adjust to individual learning preferences. This approach moves beyond traditional classroom settings by creating an adaptable and responsive learning ecosystem, tailored to the needs of each student. By leveraging AI's language generation abilities, educational content such as exercises, scenarios, and assessments can be crafted to better align with the learner's linguistic, cognitive, and cultural profile, bridging the gap between standardized curricula and the diverse needs of students.

The significance of this research lies in its potential to transform language proficiency development, not only by enhancing linguistic skills but also by fostering inclusivity, adaptability, and cultural responsiveness in English language education. By personalizing learning experiences through AI-driven, interactive learning spaces, this study aims to contribute to a paradigm shift——one that prioritizes the needs of individual learners in dynamic, culturally sensitive environments.

Building on the challenges posed by traditional English language teaching methods, this research sets forth a comprehensive set of objectives aimed at exploring the integration of Generative AI into interactive learning environments:

1. Investigating the historical context and challenges faced by traditional English language teaching methodologies, emphasizing the necessity for adaptive, culturally sensitive approaches.
2. Exploring the potential of Generative AI in crafting interactive, real-time educational content that adapts to

individual learner needs and learning contexts.

3. Evaluating the language generation capabilities of Generative AI to create personalized and contextually relevant learning experiences, including exercises and assessments.
4. Designing and implementing a novel, technology-driven oral assessment method embedded within an interactive learning space, which offers personalized feedback based on individual learner performance.
5. Assessing the effectiveness of the proposed method in enhancing student engagement, language acquisition, and overall learning outcomes across diverse linguistic and cultural contexts.

This research utilizes deep learning algorithms to create an interactive, AI-driven oral assessment system that addresses gaps in traditional language teaching methods and enhances the learning experience. Through this approach, the study aims to promote a more engaging, student-centered English education, using interactive learning tools that respond to real-time data about students' progress and needs.

The paper is structured as follows: after this introduction, the literature review section provides a detailed understanding of the historical context and existing challenges in English language education. The methodology section outlines the research design, data collection methods, and analysis techniques used to achieve the study's objectives. Following that, the results are presented and analyzed, with a thorough discussion of the findings within the context of existing research. The paper concludes with a synthesis of key insights and suggestions for future research in the integration of Generative AI into interactive learning environments.

In summary, this study aims to develop innovative, AI-driven tools that offer personalized learning experiences in English language education. By integrating interactive learning spaces and leveraging the capabilities of Generative AI, it seeks to improve both the quality of instruction and the overall student experience.

Contributions of the Paper:

- **Generative AI-Based Spoken Language Assessment:** The paper presents an AI-driven system using Transformer models to assess students' spoken English, providing personalized feedback to improve learning outcomes.
- **Diverse Spoken English Material Generation:** It introduces a method for generating diverse spoken English materials using generative models, enhancing students' learning resources and supporting comprehensive oral training.
- **Advancement of AI in Education:** The research demonstrates the potential of Generative AI in modernizing English language education by developing intelligent, interactive, and personalized learning tools and assessment systems.

The paper is organized as follows: Section II provides a review of the current research on oral assessment algorithms, offering a comprehensive understanding of existing methods and their limitations. Section III details the proposed method, including the approach for data collection, model design, and the evaluation metrics used to assess performance. In Section IV, the experimental design is discussed, encompassing the data collection process, the evaluation criteria, the experimental setup, and the comparative analyses conducted to test the proposed method. The final section concludes the paper, synthesizing key findings and suggesting potential avenues for future research in the integration of Generative AI into interactive learning environments for English language education.

2. Related Work

In this section, we provide an overview of the current research on oral assessment algorithms, focusing on their application in interactive learning environments for English language education. We begin by discussing the basic principles of oral assessment algorithms and commonly used techniques in speech evaluation. In typical oral assessment systems, the teaching process involves presenting the target text to students, followed by their attempt to read the text aloud. The system evaluates the learner's pronunciation, identifies errors, and provides feedback to

guide improvement(Zhang, Y. & Cao, J. 2022). For example, if a learner mispronounces a word like “like” (with phonemes “L, I KE”), the system detects the error and suggests corrections, thereby helping the learner achieve more accurate pronunciation and improve their speaking skills.

With advancements in speech technology, spoken language evaluation algorithms have evolved alongside the development of speech recognition models. This section first introduces key speech recognition models(Kumar, K., Kumar, R., De Boissiere, T., et al. 2019), followed by a discussion of various spoken language evaluation algorithms that have emerged as a result. For a given speech feature XXX, the task of identifying its corresponding text YYY is typically framed as an optimization problem involving maximum posterior probability. Traditional models like the Hidden Markov Model (HMM) align phonemes in the target text to HMM states, using Gaussian Mixture Models (GMM) to describe the probabilistic relationship between audio features and these states. However, with the advent of Deep Neural Networks (DNNs), which offer more powerful modeling capabilities, GMMs have gradually been replaced. As DNNs offer better recognition performance and avoid the cumbersome training processes of HMMs, researchers have focused on leveraging deep learning models for end-to-end speech recognition tasks. One notable development is the use of Connectionist Temporal Classification (CTC) loss functions, which allow the model to directly align speech features with text sequences by calculating probability sums over all possible paths(Dong, L., Xu, S. & Xu, B. 2018).

The introduction of Transformer models in natural language processing has significantly advanced speech recognition tasks. The attention mechanism, which is central to Transformer models, aligns audio features with corresponding text sequences, offering more accurate and efficient recognition than previous models like RNNs. These models, such as Listen, Attend, and Spell (LAS) and Transformer-based architectures, have been widely adopted for sequence-to-sequence tasks, including speech recognition. Transformer models, with their ability to process long-range dependencies and parallelize computations, have surpassed RNN-based models in both recognition accuracy and computational efficiency, making them an ideal framework for spoken language evaluation in interactive learning spaces.

In the context of language learning, a mispronunciation is defined as a deviation from the standard pronunciation of the target text. Early oral assessment algorithms typically compared students' pronunciation characteristics with standard reference pronunciations. These systems extracted phoneme features from both the target pronunciation and the learner's speech, using Dynamic Time Warping (DTW) to align the two sets and compute a pronunciation quality score based on the distance between corresponding phonemes. However, these comparison-based approaches are most effective for fixed assessment content, such as pre-set lessons, and become less practical when students have the flexibility to choose their own reading material(AbuSahyon, A. S. A. E., Alzyoud, A., Alshorman, O. & Al-Absi, B. 2023).

As speech recognition technology has advanced, more robust features are now extracted from audio input, leading to the development of more sophisticated spoken language evaluation algorithms. One of the most widely used algorithms is the Goodness of Pronunciation (GOP), which employs an HMM acoustic model(Hurskaya, V. 2023). The GOP algorithm assesses pronunciation quality by aligning the learner's speech features with a pre-trained standard pronunciation model, calculating the probability of phoneme correctness. Although effective, HMM-based algorithms require significant engineering efforts, such as building pronunciation dictionaries and tuning model parameters. These challenges have led to the adoption of end-to-end speech recognition models, which simplify the deployment process and provide more scalable solutions for dynamic learning environments.

End-to-end systems directly recognize phonemes from students' speech and compare them to the target pronunciation, providing feedback on mispronunciations. With the rise of deep learning and speech recognition models, these end-to-end approaches have become more accurate, enabling real-time, personalized feedback in interactive learning environments. The current state-of-the-art spoken language evaluation systems rely on deep

learning techniques to extract high-level semantic features from speech, providing more accurate and responsive feedback to learners(Zhang, Q., Lu, H., Sak, H., et al. 2020).

Interactive learning environments, powered by AI, offer exciting possibilities for oral assessment systems. These systems are designed to create dynamic, engaging spaces where learners can practice pronunciation and receive instant, personalized feedback. By integrating Generative AI into such environments, language learning systems can generate custom content—such as pronunciation exercises, scenarios, and assessments—that are tailored to each learner's proficiency level and learning style. This adaptability helps bridge the gap between standardized curricula and the diverse needs of learners, creating a more inclusive and personalized language learning experience(Won, J. & Hong, G. 2022). Furthermore, these systems enable students to engage with language content in a way that aligns with their individual learning preferences and cultural contexts, fostering a deeper connection to the material(Griffin, D. & Lim, J. 1984).

3. Method

In this section, we describe the data collection methods, model design, and evaluation analysis techniques used in this study. We begin by detailing the design of the teaching environment, including the learning context, interaction models, and scenario-based methods. Following that, we present the speaking language evaluation method based on Generative AI, which is central to the proposed approach.

3.1 Interactive Learning Environment Design

The core aim of computer-aided oral English teaching is to overcome the time and spatial limitations of traditional classroom instruction while providing immediate, personalized feedback to students. Feedback plays a pivotal role in language learning, particularly for oral English skills. Effective feedback should not only help students assess their progress but also pinpoint areas where their pronunciation deviates from the norm. Moreover, it should explain the reasons behind these errors and provide guidance on how to correct them. Unfortunately, many current oral evaluation algorithms are limited to providing only text-based feedback(Chan, W., Jaitly, N. & Le, Q. 2016), which lacks interactivity and fails to address students' specific pronunciation challenges in a meaningful way(Zhang, L., Zhao, Z., Ma, C., et al. 2020). These systems tend to offer simple, one-dimensional feedback that can be hard for learners to interpret and apply. In contrast, oral English teachers can offer richer, more intuitive forms of feedback that cater to the sensory and pronunciation needs of students(Anis, M. 2023).

To create an interactive learning environment that fosters such engagement, we design an intelligent teaching interaction model. This model involves constructing a teaching space where students are immersed in dynamic, scenario-based learning experiences. The environment responds to students ' actions in real time, offering contextualized tasks that match their learning needs. Rather than static instructional materials, the system adapts to the learner ' s input and performance(Zhao, G., Sonsaat, S., Silpachai, A., et al. 2018), providing immediate, meaningful feedback on their spoken English.

The teaching interaction model is built around a scenario-based approach, where real-world contexts and tasks are simulated. For instance, students might engage in an English conversation within a story-based scenario, where the AI system evaluates their pronunciation and fluency as part of the task. This feedback goes beyond simple error correction; it includes personalized suggestions on pronunciation improvement and contextual insights that consider the learner's progress, behavior, and the content of the lesson. By analyzing patterns of interaction between the student and the AI, we can more accurately identify areas of difficulty and offer targeted solutions.

The design process of this interactive learning environment begins with data analysis of students ' learning behaviors, focusing on their responses to various tasks and their specific areas of struggle. This analysis helps to craft a learning environment that meets the needs of each learner, using the data to generate custom scenarios that address these specific challenges. Once the interactive learning scenario model is established, it becomes easier to adapt the teaching environment dynamically, ensuring that it remains engaging, effective, and personalized. By

combining AI-driven insights with context-sensitive learning tasks(Zhang, L., Zhao, Z., Ma, C., et al. 2020), this model fosters an environment where students can interact meaningfully with the language, ensuring more effective learning outcomes(Graves, A., Fernandez, S. & Gómez, F. 2006).

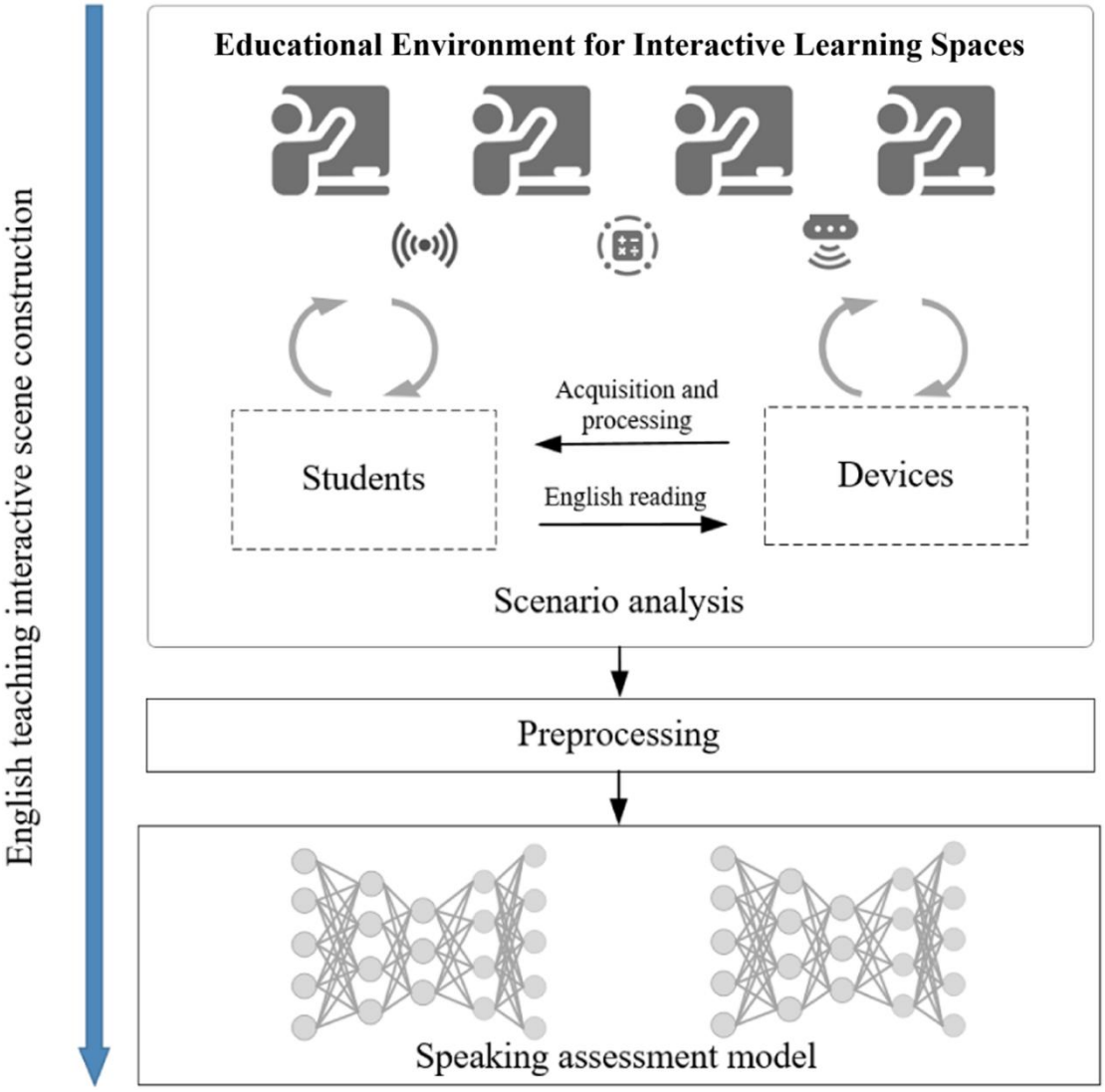

Fig. 1 Educational Environment for Interactive Learning Spaces

3.2 Speaking Assessment Model Design

Spoken language assessment consists of two main components: speech synthesis and pronunciation correction. Speech synthesis refers to the task of converting text sequences into audio feature sequences(Ren, Y., Ruan, Y., Tan, X., et al. 2019). Traditionally, models like Tacotron—which utilizes an RNN-based Encoder-RNN-Decoder structure—have been used to map text sequences to their corresponding speech signals. Tacotron aligns text with input audio using an Attention mechanism and then generates a spectrogram, which is converted into a waveform via the Griffin-Lim algorithm(Povey, D., Burget, L., Agarwal, M., et al. 2010).

However, with the advent of the Transformer model, sequence-to-sequence modeling performance has significantly improved. The Transformer model has been successfully applied to speech synthesis, and it differs from speech recognition in terms of how it handles the termination of the output sequence. While speech recognition generates discrete text sequences with sentence terminators, speech synthesis produces continuous speech signals and, therefore, also requires an additional binary classification task to predict when the output

should terminate.

While autoregressive deep speech synthesis models have outperformed traditional methods, several challenges remain. These include slow inference speed and issues with duration robustness. Due to the autoregressive nature of these models, inaccurate termination predictions may lead to incomplete or duplicated output, which impacts the synthesis quality(Leung, W. K., Liu, X. & Meng, H. 2019).

Moreover, the current speech synthesis models, such as GoogleTTS, are typically designed for single-speaker voice generation and fail to adapt to individual pronunciation needs. These systems produce a standard output based on a fixed voice, which may limit the learning experience for students. The monotonous nature of such synthesized speech does not cater to the varied and dynamic learning needs of individual students. As a result, students can only mimic the generated pronunciation without addressing their specific pronunciation errors or improving their speaking abilities(Liu, R., Chen, X. & Wen, X. 2020).

In response to these limitations, we propose a pronunciation correction model that leverages the Transformer-based architecture and integrates ideas from denoising autoencoders and speech synthesis models. The objective is to create a system that not only generates correct speech but also adapts to the learner’s spoken input, offering real-time corrections(Watanabe, S., Hori, T., Karita, S., et al. 2018).

The architecture of the proposed pronunciation correction model is shown in Fig. 2, where we combine an end-to-end automatic speech recognition (ASR) system with the Transformer. The ASR system first transcribes the spoken input into text, which is then processed by the Transformer model. To ensure transcription accuracy, we manually verify and correct the output, followed by preprocessing steps such as tokenization and formatting before feeding the data into the Transformer model for further processing and evaluation.

One of the key innovations of this model is the integration of positional encoding (PE) within the Transformer. Since the Transformer does not use an RNN structure, it explicitly incorporates positional information to account for the sequence’s temporal aspect. The positional encoding is defined as:

$$PE(pos, 2i) = \sin\left(\frac{pos}{10000^{2i/d_{model}}}\right), \quad PE(pos, 2i+1) = \cos\left(\frac{pos}{10000^{2i/d_{model}}}\right) \tag{1}$$

where pos denotes the position in the sequence, and d_{model} represents the dimensionality of the model. These positional encodings enable the model to learn the relative positions of input features effectively

The core of the Transformer architecture lies in its Multi- Head Attention (MHA) mechanism

The attention function is computed as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2}$$

where Q, K, and V are the query, key, and value matrices, respectively. This mechanism allows the model to capture relationships between different positions in the input sequence The outputs of multiple attention heads are concatenated and passed through a linear projection:

$$\text{MHA}(Q, K, V) = \text{concat}(h_1, \ldots, h_n)W^O \tag{3}$$

where each head h_i is computed as:

$$h_i = \text{Attention}(QW_Q^i, KW_K^i, VW_V^i) \tag{4}$$

For our model, nnn is the number of attention heads, set to 5. The model uses self-attention in the encoder to extract high-level features from the speech input, and cross-attention in the decoder to align the target text features with the audio features.

In the training phase, we modify the acoustic unit sequence from a standard dataset by introducing misreadings to simulate errors. For each input phoneme, the model attempts to predict the correct sequence of acoustic units based on these modified inputs. The output is a sequence that reflects corrected pronunciation. The cross-entropy loss function is used to compute the discrepancy between the predicted output and the original, canonical sequence:

$$\ddot{y}_{rec} = \text{CrossEntropy}(\ddot{\Theta}, \ddot{\Theta}) \quad (5)$$

where $\ddot{\Theta}$ is the modified acoustic unit sequence and $\ddot{\Theta}$ is the standard reference sequence By training the model in this manner, we enable it to learn both standard pronunciation and mispronunciations, allowing for real-time pronunciation correction during inference. This approach represents a shift from traditional speech synthesis models, as it directly corrects mispronunciations without requiring paired corpora or extensive engineering. By using self-supervised learning on standard datasets, the model adapts to the learner's input and corrects mispronunciations effectively.

4. Experiments

This section presents the experimental setup and evaluation of the proposed personalized English oral assessment method, integrating Generative AI within an interactive learning environment. The experiment focuses on assessing the efficacy and adaptability of our approach, considering various data sets and evaluation metrics. The experimental design includes data collection, setup details, performance metrics, and a comparative analysis with existing models to provide a comprehensive understanding of the proposed method's effectiveness in a dynamic and interactive learning space.

4.1. Implementation Details

For the implementation of our experiment, we utilized Python 3.8 as the primary programming language, with PyTorch 1.9 serving as the machine learning framework for developing, training, and deploying the Generative AI models. The models are run on a robust hardware setup consisting of an Intel Core i7–10700K CPU, an NVIDIA GeForce RTX 3080 GPU, and 32GB of DDR4 RAM, which ensures efficient processing and the ability to adapt quickly based on the data inputs. This combination of advanced hardware and software forms a high-performance, interactive environment that supports the objectives of evaluating the proposed personalized English oral assessment method.

The Transformer model used in the experiment is configured with the following parameters: 4 encoding layers, 4 decoding layers, an attention dimension of 512, a feedforward dimension of 1024, and 4 attention heads. This setup enables the model to process complex speech data and provide real-time feedback in a dynamic learning context.

4.2. Dataset and Metrics

To rigorously assess the effectiveness of the personalized English oral assessment system, we use two distinct datasets: Librispeech and L2Arctic. The Librispeech dataset offers a rich collection of over 1000 hours of English speech, primarily from native speakers, including a range of dialects and accents(Graves, A., Fernandez, S. & Gómez, F. 2006). This dataset helps evaluate the system's ability to understand and assess language comprehension and fluency in native English contexts.

The L2Arctic dataset complements this by focusing on non-native English speakers, representing a variety of first-language backgrounds. This dataset introduces challenges in assessing pronunciation, offering valuable

insights into the system's ability to adapt to diverse linguistic profiles in an interactive learning setting(Moritz, N., Hori, T. & Le, J. 2020).

For data processing, we utilize the Kaldi speech processing toolkit to extract 80-dimensional Fbank features from the audio files. We also incorporate data augmentation techniques, including both audio and text augmentation, to enrich the diversity of the spoken language dataset. Additionally, we apply transfer learning using pre-trained model parameters to improve the model's generalization ability and performance across both datasets.

Table 1 The summary of data set division on two datasets.

Dataset	Clean	Train	Val	test
Librispeech	1567	64563	2607	2183
L2Arctic	120	2500	213	895

4.3. Performance Comparison

To evaluate the performance of our proposed model, we compared it against five different models, each representing a different approach to speech recognition and synthesis. These models include HMM-ASR, a traditional phoneme-level automatic speech recognition (ASR) model based on Hidden Markov Models, TC-ASR, a text-prior-based phoneme-level recognition model, and G-ASR, the generative model proposed in this study. Additionally, we included two speech synthesis models: GoogleTTS and GlowTTS. These models were assessed using a set of metrics, with the F1 score as the primary measure of effectiveness in pronunciation correction.

The GoogleTTS and GlowTTS models, both based on speech synthesis, showed high F1 scores due to their ability to generate standard pronunciation directly from text. However, these models lack the interactive capabilities to process real-time learner input, leading to lower performance in style retention and pronunciation correction when compared to G-ASR. GlowTTS demonstrated slightly better results than GoogleTTS due to its ability to replicate the speaker's timbre, but it still struggled to preserve rhythm and accuracy, particularly in non-native speech contexts. Conversely, G-ASR showed exceptional performance across both datasets, with the highest F1 scores, by offering real-time feedback and pronunciation correction that preserved the learner’ s original speech characteristics while adapting to their unique accent and speech patterns.

In conclusion, G-ASR outperformed all other models, demonstrating its superiority in providing personalized, interactive feedback within a learning space. While traditional models like HMM-ASR and TC-ASR excelled in phoneme-level recognition, they failed to match G-ASR in terms of fluency, adaptability, and pronunciation correction in diverse linguistic contexts. The G-ASR model’ s ability to integrate real-time speech correction and maintain a learner-centered, dynamic learning environment makes it a highly effective solution for personalized English language assessment.

Table 2 The comparison results of different methods on two datasets.

Dataset	Model	F1 Score	Precision (Pre)	Recall (Rec)	Accuracy (Acc)	Specificity (spe)
LibriTTS	HMM-ASR	0.534	0.545	0.864	0.604	0.539
	TC-ASR	0.647	0.658	0.883	0.724	0.652
	G-ASR	0.735	0.761	0.945	0.815	0.748
	GoogleTTS	0.668	0.632	0.872	0.691	0.65
	GlowTTS	0.794	0.742	0.924	0.786	0.767
L2-Arctic	HMM-ASR	0.486	0.438	0.761	0.568	0.461
	TC-ASR	0.517	0.494	0.748	0.583	0.505

G-ASR	0.634	0.583	0.864	0.784	0.607
GoogleTTS	0.523	0.593	0.816	0.682	0.556
GlowTTS	0.676	0.572	0.806	0.737	0.62

4.4. Ablation Study

In order to evaluate the influence of key factors on the performance of our personalized English oral assessment method, we conducted a thorough ablation study. This study systematically examined the impact of varying parameters, including training dataset proportions, model learning rates, and model depths, to assess their effects on the model’s overall performance. By exploring these factors, we aimed to gain a deeper understanding of the model’s sensitivity to different configurations, and how these adjustments can optimize the method for use in interactive learning environments. The results offer insights into the most influential factors that drive the success of personalized language assessment systems in diverse educational settings.

(1) Training Dataset Proportions

The first part of the ablation study focused on varying the proportions of native and non-native speech data in the training datasets, specifically the Librispeech and L2Arctic datasets. We adjusted the ratio of native to non-native English speech data during training to analyze how linguistic diversity impacts the model’s ability to generalize and perform effectively in varied learning scenarios. Figures 3 and 4 illustrate the results of these experiments, showing how different training dataset proportions affect the model’s performance. The findings indicate that increasing the proportion of diverse training data results in better performance, particularly in terms of spoken language evaluation. Notably, the method achieved an F1 score above 0.6 on the L2-Arctic dataset and over 0.7 on the LibriTTS dataset when the full training set was used. Even with only 20% of the training data, the model still demonstrated strong performance, underlining the robustness and flexibility of the approach. This confirms that the model performs effectively even in scenarios where training data is limited or more varied, making it a suitable tool for dynamic, interactive learning spaces.

(2) Model Learning Rates

Another critical aspect of the ablation study was the exploration of different learning rates. Learning rate is a fundamental hyperparameter that governs the speed of model convergence and can significantly affect the model's performance. We tested a range of learning rates to identify the optimal value that balances training efficiency with accuracy. Table 3 presents the comparison of model performance under various learning rates. The results suggest that a learning rate of 0.001 produces the best balance between fast convergence and high performance. Higher learning rates tend to cause faster convergence but at the expense of precision, while lower learning rates slow down convergence but can enhance recall. This finding is important for real-time interactive learning environments where both the speed of training and the model’s ability to provide accurate feedback are crucial.

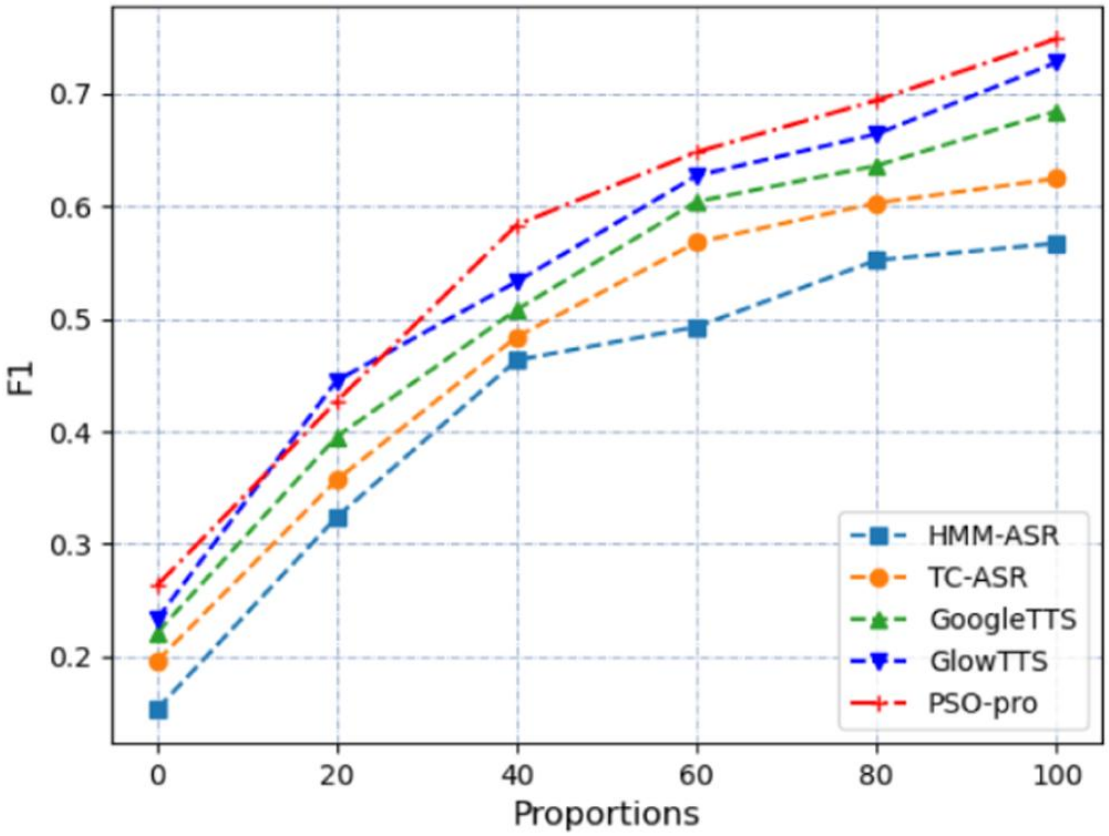

Fig. 3 The impact of different training data proportions on the F1 values for the LibriTTS dataset.

(3) Audio Correction Results and Spectrogram Comparison

To further assess the model’ s ability to maintain the semantic and rhythmic integrity of speech during correction, we visualized the spectrograms produced by the different models. Figure 5 compares the spectrograms of the original input speech (Figure 5(a)) with those generated by various models. The HMM-ASR model (Figure 5(b)) fails to preserve the rhythm and semantic information of the original input, resulting in noticeable distortion. In contrast, the TC-ASR model (Figure 5(c)) retains the original style and speech patterns more effectively. Models such as GoogleTTS (Figure 5(d)), which rely solely on text for generation, discard important speaker-specific characteristics like rhythm, leading to significant style loss. The GlowTTS model (Figure 5(e)), though capable of partially cloning the speaker’ s timbre, still cannot retain rhythm and semantic content. Finally, the model proposed in this study (Figure 5(f)) stands out by preserving both semantic meaning and rhythmic characteristics, with minimal changes to the spectrogram. This indicates the proposed method’ s superior capability in maintaining the authenticity of the original speech while providing effective feedback for improvement.

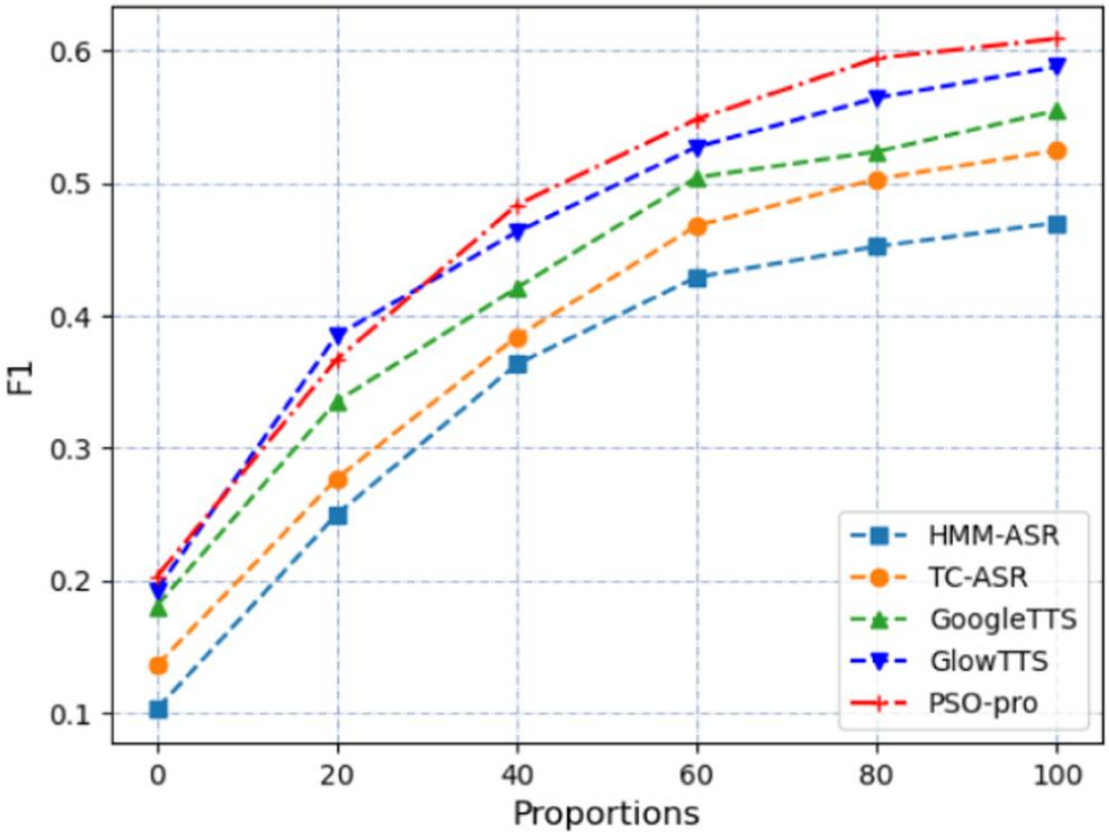

Fig. 4 The impact of different training data proportions on the F1 values for the L2-Arctic dataset.

(4) Implications for Interactive Learning Spaces

The ablation study results highlight the ability of our personalized oral assessment method to adapt to varying data conditions and to provide high-quality feedback in an interactive learning space. By fine-tuning the training dataset proportions and learning rates, we can optimize the model ' s performance for different learner profiles and language backgrounds. Furthermore, the ability of the proposed model to preserve the integrity of the learner ' s speech while providing corrections makes it especially valuable for interactive learning scenarios where real-time, personalized feedback is essential. These findings suggest that our approach is not only effective in traditional educational contexts but also highly suitable for dynamic, interactive environments where learners require constant, adaptive support to improve their English proficiency.

5. Conclusion

In conclusion, this paper presented a novel exploration into the integration of Generative AI within the context of interactive English language education. The personalized oral assessment method developed in this study represents a significant advancement in addressing the limitations of traditional teaching methodologies(Wei, L. 2023). By leveraging diverse linguistic data and focusing on real-time interaction, our approach offers a dynamic, personalized learning experience that is adaptable to the needs of each learner(Filippetti, S., Sbardella, T. & Montanucci, G. 2024). The integration of Generative AI into language learning not only enhances the educational experience but also fosters inclusivity by accounting for linguistic diversity in modern classrooms.

The findings from this research contribute to the ongoing dialogue on the role of AI in language education, presenting a pathway for the development of adaptive and culturally sensitive learning tools(Filippetti, S., Sbardella, T. & Montanucci, G. 2024). By offering personalized feedback and real-time assessment, our method supports learners at various proficiency levels, enhancing their speaking skills in interactive, real-world scenarios. This approach is particularly relevant in the context of multicultural and diverse educational environments, where

learners may come from different linguistic backgrounds(Liu, R., Chen, X. & Wen, X. 2020). The potential of this model to offer tailored feedback based on individual learning styles underscores its value in modern language education(Leung, W. K., Liu, X. & Meng, H. 2019).

Despite its strengths, the method proposed in this paper has certain limitations. The model's effectiveness is still influenced by the quality and diversity of the training data, which can be a constraint in some cases. Furthermore, challenges remain in generalizing the model's performance across various accents, dialects, and individual speech characteristics. These issues, along with the computational resources required for model training, highlight areas that need further optimization.

Looking forward, we plan to expand on this research by further refining the model and incorporating more diverse datasets to enhance its adaptability. Our future work will also focus on developing advanced personalized learning tools and assessment systems within interactive learning environments. By leveraging Generative AI, we aim to create more diverse spoken English materials and improve the accuracy of oral proficiency feedback, ultimately driving progress in English language education.

References

Baidoo-Anu, D., & Ansah, L. O. (2023). Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning. Journal of AI, 7(1), 52 – 62.

Lee, A., Chen, N. F., & Glass, J. (2016). Personalized mispronunciation detection and diagnosis based on unsupervised error pattern discovery. In Proceedings of the 2016 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) (pp. 6145-6149). IEEE.

Chimbga, B. (2023). Exploring the ethical and societal concerns of generative AI in Internet of Things (IoT) environments. In Southern African Conference for Artificial Intelligence Research (pp. 44-56). Cham: Springer Nature Switzerland.

Higuchi, Y., Watanabe, S., Chen, N., et al. (2020). Mask CTC: Non-autoregressive end-to-end ASR with CTC and mask predict. arXiv Preprint arXiv:2005.08700.

Zhang, Y., & Cao, J. (2022). Design of English teaching system using artificial intelligence. Computers and Electrical Engineering, 102, 108115.

Wang, Y., Skerry-Ryan, R. J., Stanton, D., et al. (2017). Tacotron: Towards end-to-end speech synthesis. arXiv Preprint arXiv:1703.10135.

Anis, M. (2023). Leveraging artificial intelligence for inclusive English language teaching: Strategies and implications for learner diversity. Journal of Multidisciplinary Educational Research, 12(6), 54-70.

Zhang, L., Zhao, Z., Ma, C., et al. (2020). End-to-end automatic pronunciation error detection based on improved hybrid CTC/attention architecture. Sensors, 20(7), 1809.

Graves, A., Fernandez, S., & Gómez, F. (2006). Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In Proceedings of the 23rd International Conference on Machine Learning (pp. 369 – 376).

Ren, Y., Ruan, Y., Tan, X., et al. (2019). Fastspeech: Fast, robust and controllable text to speech. Advances in Neural Information Processing Systems, 32, 1-11.

Povey, D., Burget, L., Agarwal, M., et al. (2010). Subspace Gaussian mixture models for speech recognition. In Proceedings of the 2010 IEEE International Conference on Acoustics, Speech, and Signal Processing (pp. 4330-4333). IEEE.

Leung, W. K., Liu, X., & Meng, H. (2019). CNN-RNN-CTC based end-to-end mispronunciation detection and diagnosis. In Proceedings of the ICASSP 2019—2019 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) (pp. 8132-8136). IEEE.

Liu, R., Chen, X., & Wen, X. (2020). Voice conversion with transformer network. In Proceedings of ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) (pp. 7759-7759). IEEE.

Watanabe, S., Hori, T., Karita, S., et al. (2018). Espnet: End-to-end speech processing toolkit. arXiv Preprint arXiv:1804.00015.

Zhao, G., Sonsaat, S., Silpachai, A., et al. (2018). L2-ARCTIC: Anon-native English speech corpus. Interspeech, 2783-2787.

Chan, W., Jaitly, N., & Le, Q. (2016). Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In Proceedings of the 2016 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) (pp. 4960-4964). IEEE.

Castro, G. P. B., Chiappe, A., Rodríguez, D. F. B., & Sepulveda, F. G. (2024). Harnessing AI for Education 4.0: Drivers of Personalized Learning. Electronic Journal of e-Learning, 22(5), 01-14.

Moritz, N., Hori, T., & Le, J. (2020). Streaming automatic speech recognition with the transformer model. In Proceedings of ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 6074-6078). IEEE.

Filippetti, S., Sbardella, T., & Montanucci, G. (2024, November). AI-Driven Approaches to L2 Learning: Technology Applications and Perspectives. In Conference Proceedings. Innovation in Language Learning 2024.

Lee, A., & Glass, J. (2012). A comparison-based approach to mispronunciation detection. In Proceedings of the 2012 IEEE Spoken Language Technology Workshop (SLT) (pp. 382-387). IEEE.

Panayotov, V., Chen, G., & Povey, D. (2015). Librispeech: An ASR corpus based on public domain audio books. In Proceedings of the 2015 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) (pp. 5206-5210). IEEE.

Ayeni, O. O., Al Hamad, N. M., Chisom, O. N., Osawaru, B., & Adewusi, O. E. (2024). AI in education: A review of personalized learning and educational technology. GSC Advanced Research and Reviews, 18(2), 261-271.

Wei, L. (2023). Artificial intelligence in language instruction: impact on English learning achievement, L2 motivation, and self-regulated learning. Frontiers in Psychology, 14, 1261955.

Yekollu, R. K., Bhimraj Ghuge, T., Sunil Biradar, S., Haldikar, S. V., & Farook Mohideen Abdul Kader, O. (2024, February). AI-driven personalized learning paths: Enhancing education through adaptive systems. In International Conference on Smart data intelligence (pp. 507-517). Singapore: Springer Nature Singapore.

Kumar, K., Kumar, R., De Boissiere, T., et al. (2019). Melgan: Generative adversarial networks for conditional waveform synthesis. Advances in Neural Information Processing Systems, 32, 1-11.

Yan, B. C., Wu, M. C., Hung, H. T., et al. (2020). An end-to-end mispronunciation detection system for L2 English speech leveraging novel anti-phone modeling. arXiv Preprint arXiv:2005.11950.

Dong, L., Xu, S., & Xu, B. (2018). Speech-transformer: A no-recurrence sequence-to-sequence model for speech recognition. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) (pp. 5884-5888). IEEE.

AbuSahyon, A. S. A. E., Alzyoud, A., Alshorman, O., & Al-Absi, B. (2023). AI-driven technology and Chatbots as tools for enhancing English language learning in the context of second language acquisition: a review study. International Journal of Membrane Science and Technology, 10(1), 1209-1223.

Hurskaya, V. (2023). Approaches to Personalizing the Learning Process in Teaching English with the Help of Artificial Intelligence. Академічні візії, (18).

Zhang, Q., Lu, H., Sak, H., et al. (2020). Transformer transducer: A streamable speech recognition model with transformer encoders and rnn-t loss. In Proceedings of ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 7829-7833). IEEE.

Won, J., & Hong, G. (2022). Research on smart construction education training contents using a drone simulator. Journal of Multimedia Information Systems, 9(4), 345-354.

Griffin, D., & Lim, J. (1984). Signal estimation from modified short-time Fourier transform. IEEE Transactions on Acoustics, Speech, and Signal Processing, 32(2), 236-243.

Zhang, L., Zhao, Z., Ma, C., et al. (2020). End-to-end automatic pronunciation error detection based on improved hybrid CTC/attention architecture. Sensors, 20(7), 1809.