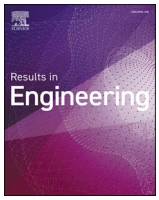
Results in Engineering 25 (2025) 103943

Contents lists available at ScienceDirect

Results in Engineering

journal homepage: www.sciencedirect.com/journal/results-in-engineering



Research paper

Automated speech therapy through personalized pronunciation correction using reinforcement learning and large language models

Ritika Lakshminarayanan [a], Ayesha Shaik [b,*], Ananthakrishnan Balasundaram [b]

[a] School of Computer Science and Engineering, Vellore Institute of Technology, Chennai 600127, India
[b] Centre for Cyber Physical Systems, Vellore Institute of Technology, Chennai 600127, India

ARTICLE INFO

Keywords:
Automatic speech recognition
Reinforcement learning
Proximal policy optimization
Large language model
Phonetic transcription
Speech synthesis markup language

ABSTRACT

Traditional approaches to pronunciation correction often face challenges in personalization, adaptability, and consistent feedback. This study introduces a novel AI-powered system that integrates Reinforcement Learning (RL) and Large Language Models (LLMs) to address these limitations. The system employs a custom Proximal Policy Optimization (PPO) algorithm for precise pronunciation evaluation and an Large Language Models to deliver detailed, encouraging, and user-specific feedback. It was evaluated using the CMU Sphinx Dictionary dataset, a foundational phonetic resource, alongside dynamically generated user-specific session data for personalized feedback and model refinement. Further validation utilized datasets such as TIMIT, LibriTTS, SpeechOcean762, and the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), enabling direct comparisons with contemporary methods. Results demonstrate the system's robustness in handling diverse phonetic variations. While primarily tested on English data, its modular architecture supports adaptation to other languages and dialects through language-specific phonetic datasets. The system achieved exceptional performance metrics: 97.9 % phoneme-level accuracy, 87.7 % word-level accuracy, 95.2 % syllable count accuracy, and 89.4 % perfect accuracy on the CMU Sphinx dataset. This innovative approach underscores the potential of advanced AI techniques to enhance the personalization and effectiveness of pronunciation correction systems. All findings are quantitatively validated and thoroughly documented.

1. Introduction

Pronunciation plays a crucial role in language proficiency, as it is essential for clear communication and understanding. Achieving correct pronunciation can be particularly challenging, requiring a solid grasp of phonetic structures, ongoing practice, and constructive feedback. When pronunciation is inaccurate, it can disrupt communication, lower confidence, and create misunderstandings—issues that are especially important for non-native speakers to address.

Despite its importance, many language learners still struggle with pronunciation, which can limit their ability to communicate effectively. This challenge is common among people of all ages, but it is particularly evident in non-native speakers who may not receive the proper guidance. This often leads to problems with pronunciation accuracy, articulation, and breaks in fluency during conversations. Research consistently shows that language-related communication struggles are a widespread issue, affecting diverse groups globally. For example, in 2010, 11.5 % of children exhibited communication difficulties, many of which were related to language challenges [1]. In the United States, approximately 7 % of children, or 1 in 14, experience developmental language disorder, which includes difficulties with phonetic aspects of language, such as phonology [2].

These challenges highlight the urgent need for accessible, personalized, and automated pronunciation correction solutions. While existing tools have evolved, they often rely heavily on manual intervention from therapists and lack a comprehensive automated approach [3]. Various machine learning (ML) techniques [7] and neural network-based methods [8,9] have been proposed for pronunciation correction, many focusing on feature extraction to identify pronunciation errors. However, these methods may not effectively capture subtle variations in pronunciation, especially in dynamic speech patterns and phonetic nuances. Furthermore, they lack the flexibility to adjust to individual learning progress, making them less effective for long-term language improvement.

Recent advancements in Automatic Speech Recognition (ASR) have significantly improved pronunciation correction. While Reinforcement

* Corresponding author.
E-mail address: ayesha.sk@vit.ac.in (A. Shaik).

https://doi.org/10.1016/j.rineng.2025.103943
Received 2 November 2024; Received in revised form 23 December 2024; Accepted 2 January 2025
Available online 3 January 2025

Learning algorithms have shown potential in enhancing speech processing tasks, challenges such as defining reward functions and interpreting decisions from deep Reinforcement Learning-based systems still persist [12,13]. The work [22] integrates Automatic Speech Recognition with phonetic analysis to provide personalized feedback for English language learners. However, it is limited by static datasets and lacks the ability to dynamically adapt to a learner's evolving pronunciation, restricting its adaptability. Similarly, the work [23] uses AI-based speech recognition and speech quality evaluation algorithms, focusing on speech clarity and intelligibility. While this method is comprehensive, it primarily evaluates speech quality rather than offering continuous, adaptive feedback, which is critical for individualized pronunciation improvement. The work [24] employs multi-modal systems using Large Language Models to assess pronunciation. While this approach offers nuanced correction, it is mainly focused on scoring accuracy and lacks iterative, personalized feedback to guide learners through their pronunciation challenges.

This paper addresses these gaps by employing a novel combination of Reinforcement Learning and Large Language Models. Unlike traditional neural network-based methods, Reinforcement Learning provides a dynamic system that tailors pronunciation correction recommendations based on real-time user input, optimizing feedback as the user progresses. Additionally, the integration of Large Language Models offers a significant advantage: while traditional methods often struggle with the naturalness and coherence of speech corrections, Large Language Models—due to their expansive linguistic understanding—can provide more contextually accurate and nuanced feedback.

This innovative combination of Reinforcement Learning and Large Language Models provides a solution to the limitations of previous approaches, particularly in the areas of adaptability and real-time correction. By leveraging Reinforcement Learning's real-time adjustment and the Large Language Model's context-aware correction, our system offers an automated, personalized approach to pronunciation correction that addresses both general language challenges and specific user needs.

The main contributions of our system are as follows:

1. Development of a personalized pronunciation correction system that integrates Automatic Speech Recognition, Reinforcement Learning, and Large Language Models.
2. Optimization of Automatic Speech Recognition for natural responses and accurate pronunciation analysis.
3. Use of Reinforcement Learning algorithms for dynamic therapy adaptation and continuous feedback improvement.
4. Integration of Large Language Models to enhance Reinforcement Learning model accuracy, providing context-rich responses.
5. Utilization of comprehensive datasets, including the CMU Sphinx Dictionary and patient session history, for enhanced model training and personalization.

The rest of the paper is organized as follows: Section 2 reviews the existing literature on pronunciation correction methodologies and technologies. Section 3 presents our innovative approach to automated pronunciation correction, detailing its underlying principles and components. Section 4 provides the outcomes of our experimental evaluations and comparative analyzes, along with performance metrics. Section 5 concludes the paper by summarizing the key findings and outlines potential areas for future work.

2. Literature review

Pronunciation improvement plays a critical role in language learning and communication proficiency.An essential step towards automating this therapy is speech recognition. To tackle this, numerous systems using technologies like automatic speech recognition, deep learning (DL), machine learning (ML), Large Language Models, and Reinforcement Learning are technologies commonly employed in such systems. However, despite their promise, existing approaches often face limitations in terms of adaptability, real-time feedback, and handling diverse linguistic contexts. This review critically analyzes the evolution of these technologies in pronunciation correction and demonstrates how our proposed approach advances the field.

Recent advancements in Automatic Speech Recognition have shown its promise for improving pronunciation learning. For example, a study conducted in southern China with 50 senior high school students demonstrated significant improvements in pronunciation accuracy, particularly for low-proficiency learners, after Automatic Speech Recognition was applied. This study also highlighted Automatic Speech Recognition's ability to better handle common pronunciation errors such as deletions and insertions [4]. Automatic Speech Recognition frameworks like Kaldi, HTK, and CMU Sphinx have become standards in language learning tools. Among these, CMU Sphinx has been noted for its strong phonetic granularity, with accuracy rates reaching 96 % in controlled environments, making it an effective tool for phoneme recognition [5,6]. However, CMU Sphinx and similar systems require model training tailored to specific languages and accents, limiting their flexibility for diverse datasets and reducing their performance when applied to non-native speech or under noisy conditions.

The SPHINX Speech Recognition System, which employs discrete hidden Markov models and LPC-derived parameters, has been particularly effective for speaker-independent and continuous speech applications. However, its accuracy still varies depending on perplexity levels, with word accuracy rates ranging from 71 % at perplexity 997 to 96 % at perplexity 20, illustrating the challenges faced in achieving high accuracy across different speech types [6]. Several studies have explored hybrid models for pronunciation assessment, particularly those combining LSTM-based RNNs with Hidden Markov Models (HMMs). These models, such as those utilizing the CMU dictionary, have achieved notable syllable count accuracy (93.8 %) and perfect accuracy (60.5 %) for phoneme-based assessment [7]. While these models show promise, their key limitation is the extended processing time required for real-time feedback, which can hinder their practicality in dynamic learning environments. Moreover, these models are limited when applied to non-native speakers or other language groups.

Another important study focused on customized datasets for Japanese learners demonstrated a Word Error Rate (WER) of just 1.97 %, underscoring the potential of language-specific datasets in improving performance [8]. However, like many other models, these systems often lack the flexibility for adaptive, real-time feedback, which is a key feature that we aim to address in our proposed system. Transformer-based models like GPT and BERT have revolutionized natural language processing (NLP) by offering powerful capabilities for generating human-like conversation and interpreting speech data. However, these models face challenges related to computational resource demands and biases in training data, limiting their real-world application [11]. Fine-tuning models such as GPT with Reinforcement Learning from Human Feedback (RLHF) has enhanced their ability to process and understand speech data, but these systems still exhibit biases that can hinder their robustness and reliability [12]. Despite this, their potential for improving Automatic Speech Recognition systems [10] and enhancing pronunciation accuracy has been demonstrated through research on datasets like Aishell-1 and LibriSpeech, where varying Word Error Rates (WER) were observed depending on the instructions given to the model [15].

To tackle the issue of context-aware feedback, multi-modal systems combining Automatic Speech Recognition with phonetic analysis have been developed. One such system achieved 90 % accuracy in phoneme-level scoring, but it lacked iterative feedback mechanisms for personalized learning [22]. Another study focusing on speech quality evaluation achieved intelligibility rates of 85 % but struggled with adaptability in dynamic learning contexts [23]. Similarly, a multi-modal Large Language Model approach that evaluated pronunciation across diverse

datasets emphasized static scoring, which limits its adaptability for real-time, user-specific corrections [24]. These systems highlight the potential of combining technologies but also reveal the necessity for dynamic, personalized feedback loops to address learner-specific challenges.

Reinforcement Learning has shown great promise in the field Reinforcement Learning language processing, particularly in systems requiring dynamic adjustments and real-time feedback. Reinforcement Learning allows systems to continuously refine their approach based on user input, making it an ideal tool for personalized pronunciation correction. The use of Proximal Policy Optimization and similar Reinforcement Learning techniques in speech systems has demonstrated improvements in speech correction accuracy by learning from user feedback [13]. However, challenges remain in defining reward functions and ensuring interpretability in Reinforcement Learning-based decision-making processes, which can introduce subjectivity and complexity in real-world applications. Semi-supervised training techniques have also been employed successfully to reduce character error rates in Automatic Speech Recognition systems, with Reinforcement Learning playing a key role in refining transcription accuracy on large datasets like the Wall Street Journal [14].

Other novel approaches to pronunciation error detection, such as anomaly detection using Deep Neural Networks (DNNs) and phoneme-specific OCSVM modeling, have achieved high accuracy across different speech types [16]. However, these systems are limited by their reliance on a narrow dataset and their inability to detect subtle errors or address non-native speaker challenges. Similarly, synthetic speech generation methods such as P2P and S2S conversion have outperformed traditional models in detecting pronunciation errors but face generalization issues and limitations in multilingual support [17]. These studies highlight the need for more adaptable systems that can handle diverse speech types and learner profiles.

The importance of trust in Artificial intelligence-based systems, particularly in healthcare and educational contexts, cannot be overstated. Research has shown that trust dynamics, influenced by factors like attachment security and anxiety, play a significant role in user interaction with Artificial intelligence systems [18]. Systems that promote trust through transparency and empathetic interaction can enhance user engagement and learning outcomes. In healthcare, trust is further influenced by the system's ability to offer reliable, transparent, and accountable recommendations, a critical component when designing systems aimed at supporting personalized learning and therapy [19].

Despite the advancements in Automatic Speech Recognition, and pronunciation correction systems, significant challenges remain in terms of adaptability, scalability, and real-time feedback. Our system builds upon previous work by integrating Reinforcement Learning with Large Language Models to provide personalized, dynamic, and context-aware pronunciation corrections. Unlike previous systems, which often rely on static scoring or fixed datasets, our approach offers a scalable solution capable of adapting to diverse linguistic challenges and providing real-time corrections tailored to the user's unique needs. The following section outlines our proposed methodology, which integrates these technologies into a unified system for automated pronunciation correction Table 1.

Table 1
Summary of key features and limitations of reviewed studies in pronunciation correction.

Refs. Number	Key Features	Limitations
[4]	Demonstrated significant improvement in pronunciation accuracy for low-proficiency learners.	Focused on a specific demographic; lacked scalability to diverse linguistic groups.
[5]	High phonetic granularity with accuracy rates reaching 96 % in controlled environments.	Requires bespoke training tailored to specific languages and accents, limiting flexibility.
[6]	Achieved word accuracy rates of 71 % at perplexity 997 and 96 % at perplexity 20.	Struggled to maintain accuracy across varying speech complexities and noisy environments.
[7]	Achieved 93.8 % syllable count accuracy and 60.5 % perfect phoneme-level accuracy.	Processing times were extended, limiting real-time feedback; less effective for non-native speakers.
[8]	Extremely low Word Error Rate (WER) of 1.97 %, highlighting benefits of customized datasets.	Lacked real-time adaptability; not generalizable to other languages or broader populations.
[11]	Revolutionized NLP by enabling powerful speech data interpretation capabilities.	Computationally demanding; prone to data biases in training, limiting robustness in real-world contexts.
[12]	Improved Automatic Speech Recognition systems and enhanced pronunciation accuracy using Reinforcement Learning from Human Feedback (RLHF).	Sensitivity to biases introduced during training; scalability challenges due to high resource demands.
[15]	Demonstrated varying Word Error Rates (WER) based on model instructions, showcasing adaptability of GPT models.	WER variability highlights sensitivity to training instructions; lacks robustness across diverse datasets.
[22]	Combined Automatic Speech Recognition and phonetic analysis to achieve up to 90 % accuracy in phoneme-level scoring.	Lacked iterative and personalized feedback, limiting dynamic adaptability.
[23]	Achieved intelligibility rates of 85 %, emphasizing clear speech processing.	Static feedback methods constrained adaptability in dynamic learning environments.
[24]	Evaluated pronunciation across diverse datasets, emphasizing accuracy.	Focused on static scoring, with limited real-time personalization and iterative feedback.
[13]	Enhanced real-time feedback and corrections using Reinforcement Learning.	Challenges in defining reward functions and interpretability in real-world settings.
[14]	Reduced character error rates significantly on large datasets like the Wall Street Journal.	Sparse user feedback and subjective evaluations introduced complexity in refining results.
[16]	Achieved high accuracy across various speech types through phoneme-specific modeling.	Relied on narrow datasets, limiting subtle error detection and adaptability for non-native speakers.
[17]	Outperformed traditional models in detecting pronunciation errors using synthetic speech generation.	Generalization issues and lack of multilingual support constrained broader application.
[18]	Explored the role of trust in user interactions, emphasizing transparency and empathetic interaction.	Required careful balancing to avoid loss of user confidence or trust issues in sensitive applications.
[19]	Focused on trust-building in healthcare AI through reliability and accountability.	Limited discussion on implementation strategies for scaling to diverse populations.

3. Methodology

3.1. Dataset

For this research, we utilized two primary datasets to facilitate the training and evaluation of our proposed methodologies. First, the CMU Sphinx Dictionary, an open-source and machine-readable dataset, served as the foundation for training both the Reinforcement Learning algorithm and the Large Language Model. This dataset, comprising over 134,000 words and their corresponding pronunciations, is invaluable due to its comprehensive coverage of linguistic elements. The CMU Sphinx Dictionary uses the CMUBET format, a variant of the ARPABET phoneme system, ensuring compatibility with our computational frameworks.

In addition to using the CMU Sphinx Dictionary dataset, we leveraged historical therapy session data dynamically generated and stored by the system. This dataset is unique and collected directly during

therapy sessions, providing a personalized source of information. It includes the following components:

- Target Word: The word the system prompts the user to pronounce.
- Correct Phonetic Transcription: The correct pronunciation of the target word, stored in IPA format.
- User's Phonetic Transcription: The user's attempted pronunciation of the target word, stored in IPA format.
- Fig. 1 illustrates an example of how user data is stored in the system during a session.

These data points are integral to refining and personalizing the models based on individual user performance over time. By incorporating these historical user interactions and feedback, our approach addresses the diverse and specific needs of each participant. The system is designed to be simple and intuitive, targeting a broad demographic without focusing on a specific age group. Through the integration of these datasets, our research offers a robust and tailored solution for personalized pronunciation correction.

To optimize model training and ensure comprehensive pronunciation coverage, we carefully curated a supplementary dataset consisting of approximately 50,000 words from the CMU Sphinx dictionary. This dataset was designed to encompass the full range of English phonemes, including various sound types like fricatives, stops, affricates, and more.

The selection criteria for these words included:

1. Phoneme Diversity: Ensuring the dataset covers all phonemes in the English language, providing a broad spectrum of sounds necessary for robust model training.
2. Frequency of Use: Prioritizing commonly used words to enhance the model's practical utility and relevance in everyday speech.
3. Phonetic Complexity: Including words with varied phonetic structures to challenge the model and improve its ability to handle complex pronunciations.
4. Varying Pronunciations: Incorporating words that are likely to have multiple pronunciations to improve the model's flexibility and adaptability in recognizing different speech patterns.

Meticulous selection ensured that the dataset reflected the diverse phonetic landscape of the language. Furthermore, to accommodate the distinct requirements of different system modules, our preprocessing stage included the conversion of CMUBET formatted words into the International Phonetic Alphabet (IPA) format. This conversion step established consistency and seamless integration across our system's components.

3.2. System architecture

In this work, we propose a novel approach to automated personalized pronunciation correction leveraging advanced technologies such as Natural Language Processing, Reinforcement Learning and Large Language Models in order to dynamically adapt therapy recommendations based on user input for optimized pronunciation feedback. Additionally, we incorporate Large Language Models to enhance the naturalness, coherence, and accuracy of the Reinforcement Learning model. The proposed system architecture is designed to be modular, offering flexibility and scalability in addressing various aspects of personalized pronunciation correction, as shown in Fig. 2. Comprising four core modules, the architecture includes:

1. Speech-to-Text
2. Reinforcement Learning
3. Large Language Model
4. Text-to-Speech

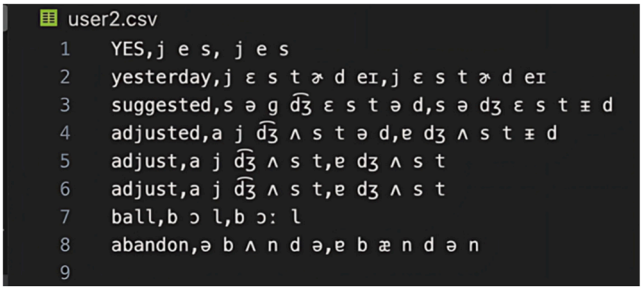


Fig. 1. Stored session data of sample user.

3.2.1. Speech-to-text module

This module transcribes user speech into phonetic text, enabling precise analysis of pronunciation nuances. Recognizing the limitations of basic Speech-to-Text (STT) toolkits, which often lack the capability to transcribe speech into phonetics directly, we have implemented an advanced approach. Our system uses Allosaurus, a pre-trained universal phone recognizer, to perform allophone-based transcription, ensuring detailed phonetic analysis and accuracy.

A major challenge with conventional STT models, such as the Google Speech-to-Text API [25], DeepSpeech [26], and WhisperASR [27], is their focus on generating standard word-level transcriptions rather than phonetic representations. These models excel at converting speech to text but fall short when it comes to capturing the nuances of phonemes. For example, while the Google Speech-to-Text API and DeepSpeech transcribe the spoken words accurately, they cannot capture phonetic discrepancies, which are essential for identifying mispronunciations. This limitation is particularly critical in our system, where accurate phonetic transcription is necessary for effective feedback.In contrast, Allosaurus [20] provides phonetic-level transcription, directly converting audio input into phonemes as pronounced, without any corrections. This capability allows our system to pinpoint which phonemes are mispronounced and deliver precise feedback based on phonetic details. By directly transcribing the audio input into its exact phonetic representation, Allosaurus enables a more detailed and accurate analysis of pronunciation errors.

Moreover, Allosaurus supports over 2000 languages, making it highly adaptable to diverse linguistic contexts. This versatility ensures that our system can be easily expanded to accommodate different languages by simply training it on the relevant datasets, allowing for global scalability. The user's audio input is captured in a WAV file to preserve its original quality. Afterward, the audio is processed using Google Speech-to-Text for basic word recognition, even if mispronounced. The recognized word is then passed through the Phonemizer plugin to convert it into International Phonetic Alphabet (IPA) transcription. Simultaneously, the audio is analyzed by Allosaurus, which generates the phonetic transcription, outputting IPA representations based on the user's exact pronunciation. These dual transcriptions are then forwarded to the Reinforcement Learning and Large Language Model modules for further analysis, enabling the system to detect mispronounced phonemes and provide tailored feedback.

3.2.2. Reinforcement learning module

This module plays a pivotal role in our system, acting as the core decision engine. It leverages a custom Proximal Policy Optimization algorithm trained on the CMU Sphinx Dataset. Unlike traditional approaches, Reinforcement Learning operates through an iterative process of experimentation and reward system refinement. Here, we delve into the unique aspects of our implementation tailored for pronunciation correction. Proximal Policy Optimization is a policy gradient algorithm commonly used in Reinforcement Learning tasks. It aims to maximize the expected reward (R) received by the agent (our Reinforcement Learning module) over a given trajectory (a sequence of actions taken by

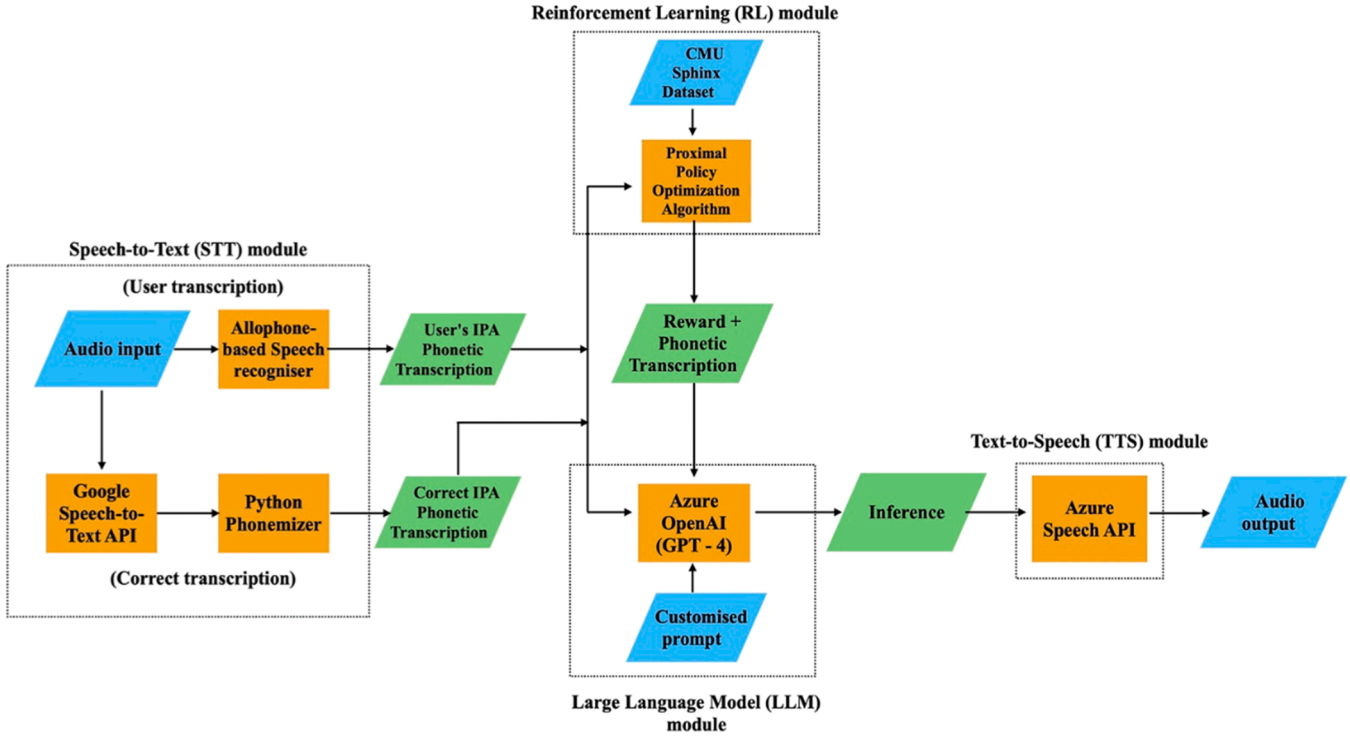


Fig. 2. Proposed system architecture.

the agent) is given in Eq. (1). This can be mathematically expressed as:

$$J(\theta) = E\pi_\theta[\Sigma\gamma^t \times R(s_t, a_t)] \quad (1)$$

where $E\pi_\theta$ denotes the expectation over trajectories generated by the policy π_θ, $J(\theta)$ represents the expected reward for a policy parameterized by θ (policy parameters of the Reinforcement Learning agent), γ is the discount factor (weighing the importance of future rewards), $R(s_t, a_t)$ represents the reward obtained at state s_t by taking action a_t and t signifies the time step within the trajectory. The algorithm achieves this objective by employing a clipping mechanism that restricts the policy update during training, and can be expressed mathematically as given in Eq. (2):

$$\pi'(a_t|s_t) = clip\left(\pi_{\theta_{old}}(a_t|s_t) \,/\, \pi_\theta(a_t|s_t), ratio_l, ratio_h\right) \times A_t(s_t, a_t) \quad (2)$$

where $\pi'(a_t|s_t)$ represents the probability of taking action a_t in state s_t under the new policy, $\pi_{\theta_{old}}(a_t|s_t)$ denotes the probability of taking action a_t in state s_t under the old policy, $clip(\ldots, ratio_l, ratio_h)$ is the clipping function that restricts the policy update ratio within a predefined range, $A_t(s_t, a_t)$ represents the advantage function, which estimates the advantage of taking action a_t in state s_t compared to the average action. This can be expressed as given in Eq. (3):

$$A_t(s_t, a_t) = Q(s_t, a_t) - V(s_t) \quad (3)$$

where $Q(s_t, a_t)$ is the Q-value function, estimating the expected future rewards of taking action $'a_t'$ in state $'s_t'$ and $V(s_t)$ is the state-value function, estimating the expected future rewards from state $'s_t'$

3.2.3. Reward shaping

An effective reward function is crucial for guiding the agent's learning process in the pronunciation correction task. For our system, the reward function has been carefully defined such that it considers both phoneme-level and word-level accuracy. The reward function, R, is defined as given in Eq. (4):

$$R = C_p - \left(T_p - C_p\right) \quad (4)$$

where C_p represents the number of correctly pronounced phonemes in the word and T_p represents the total number of phonemes in the word. This function incentivizes the agent to focus on both fine-grained phoneme accuracy and overall word pronunciation. If all phonemes are pronounced correctly $(C_p = T_p)$, the reward is maximized. However, if there are errors $(C_p < T_p)$, the reward remains positive, encouraging the agent to attempt corrections and potentially earn a higher reward in subsequent steps.

3.2.4. Hyperparameter tuning

The effectiveness of the algorithm hinges on carefully chosen hyperparameters. We conducted a rigorous hyperparameter tuning process to optimize the learning rate (α), discount factor (γ), and clipping parameter (ϵ) within the Proximal Policy Optimization framework.

- Learning Rate (α) controls the magnitude of updates to the actor and critic networks during training. The learning process may become unstable if the learning rate is set too high, while convergence may be slowed down if it is set too low. We tuned α to achieve a balance between these extremes, ensuring efficient learning within the pronunciation correction domain.
- Discount Factor (γ) determines the importance of future rewards compared to immediate rewards. A high γ prioritizes long-term goals, whereas a low γ focuses on immediate rewards. In our implementation, γ was carefully tuned to strike a balance between immediate feedback on pronunciation attempts and the long-term objective of accurate word production.
- Clipping Parameter (ϵ) is used in the algorithm's clipping mechanism to maintain stability during policy updates. It restricts the policy update ratio within a predefined range. We tuned ϵ to prevent excessive policy updates that could destabilize the learning process, ensuring smooth policy improvements while addressing pronunciation errors.

After extensive tuning to balance exploration and exploitation while ensuring stability in policy updates, the corresponding parameter values

are presented in Table 2.

3.2.5. Environment design

The pronunciation environment serves as a controlled training ground for the agent. It mimics real-world speech scenarios and incorporates mechanisms to introduce variations, fostering robust learning.

- Action Space: The current design does not utilize an explicit action space. This is because the focus is on evaluating pronunciations rather than the agent directly controlling speech production. The single action performed by the algorithm involves using the correct and user transcripts to provide a reward based on the number of correctly pronounced phonemes. Through exploration and exploitation, it tries to maximize the reward within the next 1–2 episodes for that word.
- Mistake Introduction: The environment has the capability to introduce errors in the words with a specified probability during the training process. This has been incorporated to ensure the agent is capable of handling a variety of pronunciations, and pinpoint accurately if a pronunciation is wrong in real-time.

By incorporating such detailing in the environment, we ensure that it mimics a realistic foundation upon which the Reinforcement Learning agent can be trained, thus allowing it to effectively correct the user's pronunciation.

3.2.6. Neural network architecture

The custom Proximal Policy Optimization (PPO) implementation utilizes feedforward neural networks to represent the actor and critic components. These networks process the state representation derived from the user's pronunciation attempt and guide the Reinforcement Learning agent's decision-making. Here's a breakdown of the network structure, which is also specified in Table 3:

- Input Layer: The input layer's size corresponds to the dimensionality of the state representation derived from the Speech-to-Text output and any other relevant environmental features.
- Hidden Layers: Our network includes three hidden layers with 128, 128, and 64 units, respectively. Each hidden layer utilizes ReLU activation functions to introduce non-linearity, crucial for learning complex patterns in pronunciation data.
- Output Layer:
  - Critic Network: The critic network's output layer consists of a single linear unit, producing the state-value estimation – how "good" the current state is for the agent. The critic network learns to estimate the state-value by updating its parameters to minimize the difference between its prediction and the actual observed returns. This update process can be represented as given in Eq. (5):

$$Q(s_t, a_t) = Q(s_t, a_t) + \alpha[R_{t+1} + \gamma V(s_{t+1}) - Q(s_t, a_t)] \quad (5)$$

where:

- $Q(s_t, a_t)$ is the estimated future rewards for taking action $'a_t'$ in state $'s_t'$.

Table 2
Hyperparameter values for the custom proximal policy optimization algorithm.

Parameter	Value
Learning Rate (α)	0.0003
Discount Factor (γ)	0.99
Clipping Parameter (ϵ)	0.2

Table 3
Neural network specifications for the reinforcement learning module.

Layer Type	Number of Units/ Neurons	Activation Function
Hidden Layer 1	128	ReLU
Hidden Layer 2	128	ReLU
Hidden Layer 3	64	ReLU
Output (Critic)	1	Linear
Output (Actor)	1	–

- α is the critic network's learning rate, controlling the magnitude of updates.
- R_{t+1} refers to the reward received after taking action $'a_t'$ in state $'s_t'$.
- γ is the discount factor, prioritizing immediate rewards (low γ) or long-term goals (high γ).
- $V(s_{t+1})$: The estimated future rewards from the next state $'s_{t+1}'$.
- Actor Network: The actor network's output layer size is determined by the action space. Since the current design focuses on pronunciation evaluation without direct control, it produces the current transcription for that episode and the episodic reward based on the number of correctly pronounced phonemes.

Algorithm 1 describes the logic behind the custom Proximal Policy Optimization algorithm used for our application.

When compared to typical Reinforcement Learning algorithms, this custom Proximal Policy Optimization algorithm performs one action: using the correct and user transcripts to provide a reward based on the number of correctly pronounced phonemes.

The correct transcription from the Speech-to-Text module is passed into the Reinforcement Learning algorithm's environment, whereas the user-generated transcript is provided to the actor network. The actor network's role is to maximize the reward by eliminating one phonetic error per episode in the transcript. By learning from the user's mispronunciations, the actor network refines the model's predictions over time, improving the accuracy of the system's phonetic correction. Based on this actions it performs (corrections made to the wrong phonemes), the critic network provides feedback.

3.2.7. Large language model module

Recent research highlights the importance of trust in AI systems. How transparent an AI system is about its processes directly influences user trust levels, impacting its overall success. Building the right level of trust is particularly critical for AI-powered pronunciation correction. Users should feel confident in the system's capabilities while understanding its limitations. This fosters a positive learning experience where the system serves as a helpful tool [21].

This module, powered by cutting-edge technology like GPT-4, plays a pivotal role in refining the Reinforcement Learning module's output and enhancing the overall quality of the system's responses. The corrected transcript from the Reinforcement Learning module, the final reward generated during the Reinforcement Learning process, and the user transcript identified by the Speech-to-Text module (Allosaurus) are all provided as inputs to the Large Language Model module.

This module's responsibility is to analyze the Reinforcement Learning module output and the user transcript to infer inaccuracies in the transcription and rectify them. By leveraging its advanced language processing capabilities, the Large Language Model analyzes the phonetic input, considering not only individual sounds but also broader context and linguistic nuances.

The Large Language Model ensures that the corrections are not just accurate at the phoneme level, but also contextually appropriate based on surrounding words and sentence structures. This context-aware feedback helps make the pronunciation correction more natural and user-friendly. Additionally, the Large Language Model provides positive and encouraging feedback to foster user trust, teaching them where they went wrong and offering guidance on how to improve. This analyzed

Algorithm 1
Proximal Policy Optimization (PPO) for Pronunciation Improvement.

Input: Actor network parameters θ, Critic network parameters σ, Maximum iterations M, Trajectory length T, Clipping threshold ϵ, Discount factor γ, Learning rate α
Output: Optimized actor network parameters θ, Optimized critic network parameters σ

1: Initialize θ and σ
2: for $i = 1$ to M iterations:
3: Collect trajectories by running policy π_θ for T timesteps.
4: a. Feed the user transcript into the actor network.
5: b. Feed the correct transcript into the environment.
6: c. Evaluate pronunciation and assign a reward based on number of correct phonemes.
7: d. Estimate advantages At using the reward function $R = C_p - (T_p - C_p)$
8: e. Update old policy $\pi_{old} \leftarrow \pi_\theta$.
9: for $j = 1$ to N iterations:
10: Update θ using the policy gradient method:
11: Compute clipped policy gradient with respect to θ for stability:
$$\pi'(a_t|s_t) = clip\left(\pi_{\theta_{old}}(a_t|s_t)/\pi_\theta(a_t|s_t), ratio_l, ratio_h\right) \times A_t(s_t, a_t)$$
12: Update σ:
13: Update the state-value function $Q(s_t, a_t)$ using the temporal difference error:
$$Q(s_t, a_t) = Q(s_t, a_t) + \alpha[R_{t+1} + \gamma V(s_{t+1}) - Q(s_t, a_t)]$$
14: end for
15: end for

output is then passed on to the Text-to-Speech module, which generates the final audio feedback with the corrected pronunciation.

To tailor the Large Language Model's analysis and provide more nuanced corrections, we employ a two-pronged approach to fine-tuning GPT-4. First, we leverage the custom 50,000-word dataset described earlier, which is specifically curated from the CMU Sphinx dictionary. This domain-specific data supplements GPT-4′s general training data, enabling it to better understand the intricacies of phonetics and common pronunciation mistakes. Second, we utilize specialised prompts designed to focus the Large Language Model's analysis on identifying and correcting errors within the phonetic transcript. These prompts guide the Large Language Model towards a more targeted evaluation, going beyond simply generating alternative phrasings. By combining the custom dataset and specialised prompts, we achieve a more effective and contextually aware fine-tuning process for GPT-4 within the domain.

Trust and user experience are central to the effectiveness of the proposed pronunciation correction system, particularly in speech therapy, where users rely on consistent and empathetic feedback. The Large Language Model module is fine-tuned to deliver corrections in a positive tone, recognising user successes while constructively addressing errors. Sentiment analysis ensures that responses remain professional yet empathetic, fostering a supportive learning environment. Also, the system provides feedback to individual performance trends based on user-specific session data. This sense of personalization enhances the confidence of user over time. Additionally, the user friendly interface promotes the system to be easily accessible regardless of technical expertise and age.

To ensure effective and personalized feedback, the Large Language Model uses a carefully crafted prompt that integrates several key features such as user engagement and positivity, phonetic focus, instant instructional guidance, contextual analysis and interactive learning. By incorporating these themes, the Large Language Model module not only corrects pronunciation errors but also builds user trust and confidence, creating a positive and effective learning experience. The resulting output from the Large Language Model is then fed into the Text-to-Speech module, ensuring a seamless integration of feedback and further practice. The deployment of AI-powered systems in education and therapy demands a focus on ethical considerations to ensure their responsible and fair usage. Our automated pronunciation correction system addresses several critical concerns, including accessibility, data privacy, trust and transparency.

3.2.8. Text-to-speech module

The final component of the system architecture is the Text-to-Speech module, responsible for converting the system's text-based output into audible speech for the user. The output from the Large Language Model, which is trained using the customized prompt to generate natural-sounding human-like text, is fed into the module.

To ensure accurate phonetic pronunciation and natural-sounding speech, our system leverages Speech Synthesis Markup Language (SSML) in conjunction with the Azure Speech API. Unlike basic Text-to-Speech algorithms that lack support for SSML tags, the Azure Speech API offers comprehensive functionality and diverse options for selecting natural-sounding voices. By incorporating SSML tags, we can control various aspects of pronunciation, such as pitch, emphasis, and speaking rate, ensuring clear and accurate delivery of the system's feedback and instructions.

This module plays a critical role in establishing a user-friendly and interactive communication channel, enabling the system to provide clear feedback and instructions to the user throughout the therapy session.

4. Experimental results and discussion

4.1. Evaluation metrics

The metrics used for evaluating the model are discussed below.

4.1.1. Word-level accuracy

Word-level accuracy provides a fundamental measure of the system's ability to correctly pronounce complete words. This metric offers a straightforward assessment of overall pronunciation success. It's particularly valuable for tracking progress over time and identifying broad areas where the user may need additional practice. It is calculated as given in Eq. (6):

$$WLA = (w_c / w_t) * 100\% \tag{6}$$

where WLA = Word-level Accuracy, w_c = Number of correctly pronounced words and w_t = Total number of words

4.1.2. Phoneme-level accuracy

Phoneme-level accuracy focuses on the accuracy of pronunciation of individual phonemes. This metric is particularly useful to understand the specific pronunciation challenges the user faces and personalize their therapy session accordingly. For example, a user might have high word-level accuracy but struggle with particular consonant blends or vowel sounds, which phoneme-level analysis would reveal. It is calculated as given in Eq. (7):

$$PLA = (p_c / p_t) * 100\% \tag{7}$$

where, PLA = Phoneme-level Accuracy, p_c = Number of correctly pronounced phonemes and p_t = Total number of phonemes.

4.1.3. Syllable count accuracy

Syllable Count Accuracy [7] measures the precision of predicting the number of syllables in words, particularly beneficial for words not commonly found in dictionaries as given in Eq. (8):

$$SCA = 1 - |PSC - ASC|/Max(ASC, 1) \tag{8}$$

where, SCA = Syllable Count Accuracy, PSC = Predicted Syllable Count and ASC = Actual Syllable Count.

4.1.4. Perfect accuracy

Perfect Accuracy [7] measures the precision of predicting the number of perfect predictions with respect to the total number of words as given in Eq. (9):

$$PA = n_p/n_t * 100\% \tag{9}$$

where, PA = Perfect Accuracy, n_p = Number of perfect predictions and n_t = Total number of predictions

4.1.5. Sentiment scores

Sentiment scores provide insights into the encouraging and supportive nature of the Large Language Model's responses during the sessions. These scores are typically generated using sentiment analysis tools that analyze the word choice and linguistic patterns within the Large Language Model's feedback. By tracking the sentiment expressed in the responses, we can assess how effectively it balances constructive criticism with positive reinforcement. This analysis is crucial for optimizing the Large Language Model's feedback style, ensuring it maintains an encouraging tone even when highlighting areas for improvement. A positive and supportive approach can significantly enhance user engagement and motivation during therapy sessions. Sentiment scores are usually generated by specialized tools that analyze text or speech. The emotions detected by the tools are typically classified as positive, negative and neutral, based on key words, phrases and patterns in the generated response.

4.2. Experimental results and performance evaluation

The model's performance has been tested on the remaining words from the CMU Sphinx dataset that were not used for training the model, as well as on custom user input provided while interacting with the system. The Reinforcement Learning module is the core decision engine of the system, and Fig. 3 illustrates the reward per episode generated by the reward function during Proximal Policy Optimization model training on the CMU Sphinx text dataset for 110 episodes. Red dots signify episodes where phonemes were initially incorrect and subsequently rectified by the Proximal Policy Optimization model, represented by the green dots. Since the model is tuned to work on an input till it maximizes the reward for it before moving on to the next input, it maximizes the accuracy of the model's output.

As depicted by Fig. 4, the graph shows a steady increase in the cumulative rewards over all the episodes without any sharp deviations from one episode to another, demonstrating the model's ability to learn and improve pronunciation accuracy in small steps so as to aid the user's therapy session.

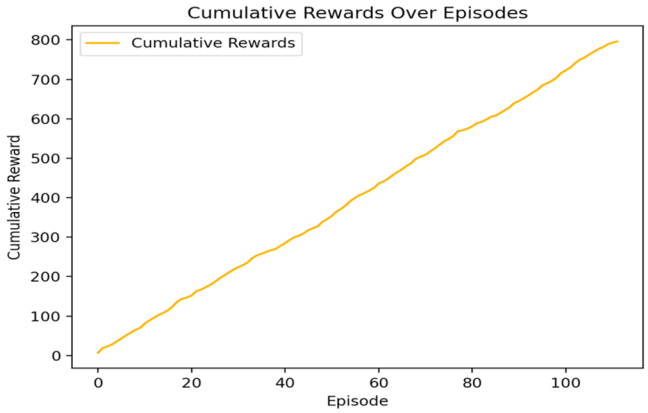


Fig. 4. Cumulative rewards generated by the Reinforcement Learning model for 110 episodes.

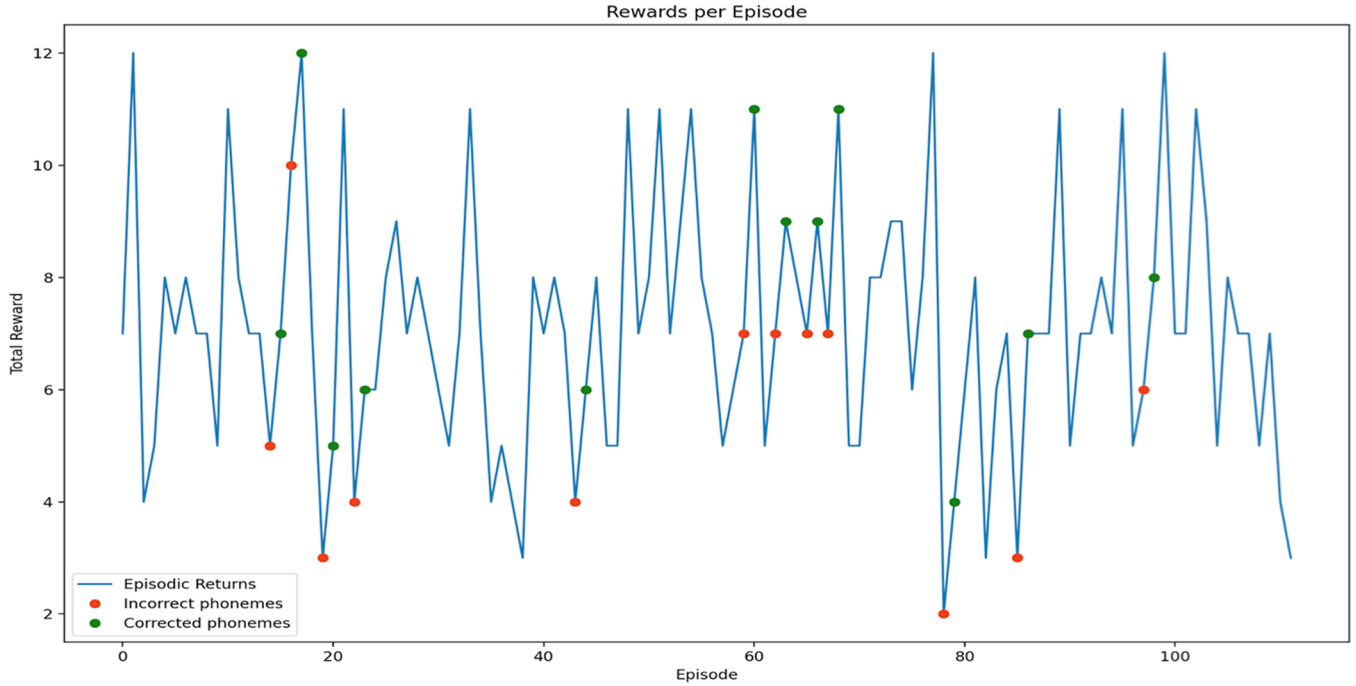


Fig. 3. Reward generated per episode by the Reinforcement Learning model for the CMU Sphinx test dataset.

Here, 'small steps' refer to the incremental adjustments in the policy and feedback mechanisms during training. Instead of making abrupt changes to its decision-making process, the model refines its learning iteratively, ensuring stable improvements in performance while minimizing errors. This approach aligns with the principles of Proximal Policy Optimization, where policy updates are constrained to avoid drastic shifts, promoting stable and efficient learning.

If the changes were drastic, the balance between exploration and exploitation would be disrupted, potentially causing the policy to oscillate between suboptimal strategies. This could lead to larger fluctuations in the cumulative reward range, requiring significantly more episodes to stabilize and converge on accurate phoneme transcriptions. By maintaining small, controlled updates, the model avoids this instability and achieves efficient convergence, allowing it to deliver reliable corrections within fewer episodes.

In addition, the performance of the model has been measured using Phoneme - Level accuracy and Word - Level accuracy metrics. As seen in Fig. 5, the model achieves an average phoneme-level accuracy of 97.5 % and an average word-level accuracy of 84.7 % respectively for 110 episodes, as depicted by the red lines in the graphs, which shows high performance quality of the model.A detailed analysis of the accuracy levels of the Reinforcement Learning model for varying number of episodes has been mentioned in Table 4. Overall, these results exhibit that the model is a highly effective and efficient method for pronunciation correction.

Table 4 shows a slight decline in both phoneme-level and word-level accuracies as the number of episodes increases. Initially, high accuracies at 50 episodes suggest quick learning of basic patterns. However, with more episodes, the model encounters diverse and potentially ambiguous examples, introducing noise and causing minor fluctuations. The Reinforcement Learning paradigm involves exploration and exploitation strategies. Early episodes are more exploratory, leading to rapid performance improvement. As training progresses, the balance between exploration and exploitation can temporarily affect accuracy. Additionally, prolonged training can lead to overfitting on specific pronunciation nuances, reducing generalization capability.

Table 4
Accuracy levels of Reinforcement Learning model on varying number of episodes.

Number of episodes	Phoneme-Level Accuracy	Word-Level Accuracy
50	0.982	0.909
100	0.974	0.872
200	0.976	0.881
500	0.972	0.868
1000	0.958	0.850

In this context, the stopping criteria for training were based on the observed trends in cumulative rewards (Figs. 3 and 4) and the accuracy metrics shown here. Training was halted at 500 episodes as the model had reached a reward plateau, and further training introduced a slight decline in accuracy metrics. This decision reflects a trade-off between achieving optimal accuracy and avoiding overfitting or computational inefficiencies. Preliminary experiments confirmed that additional training beyond 500 episodes did not yield meaningful improvements in system performance. The results at this stopping point demonstrate robust phoneme- and word-level accuracies while ensuring computational efficiency.

Several other models exist to facilitate computer-assisted pronunciation training (CAPT) using phonemes. One such model is a LSTM-RNN+HMM based model, which has been trained on the CMU Sphinx dataset. The results of this model have been compared with our model as depicted in Fig. 6.

To validate the accuracy and robustness of our system across varied user inputs, we tested the model on multiple datasets with diverse linguistic contexts, including TIMIT, LibriTTS, SpeechOcean762, and Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). These datasets represent a wide range of linguistic variations, including non-native English accents with varying levels of phonetic complexity.These datasets also allowed us to compare our model's

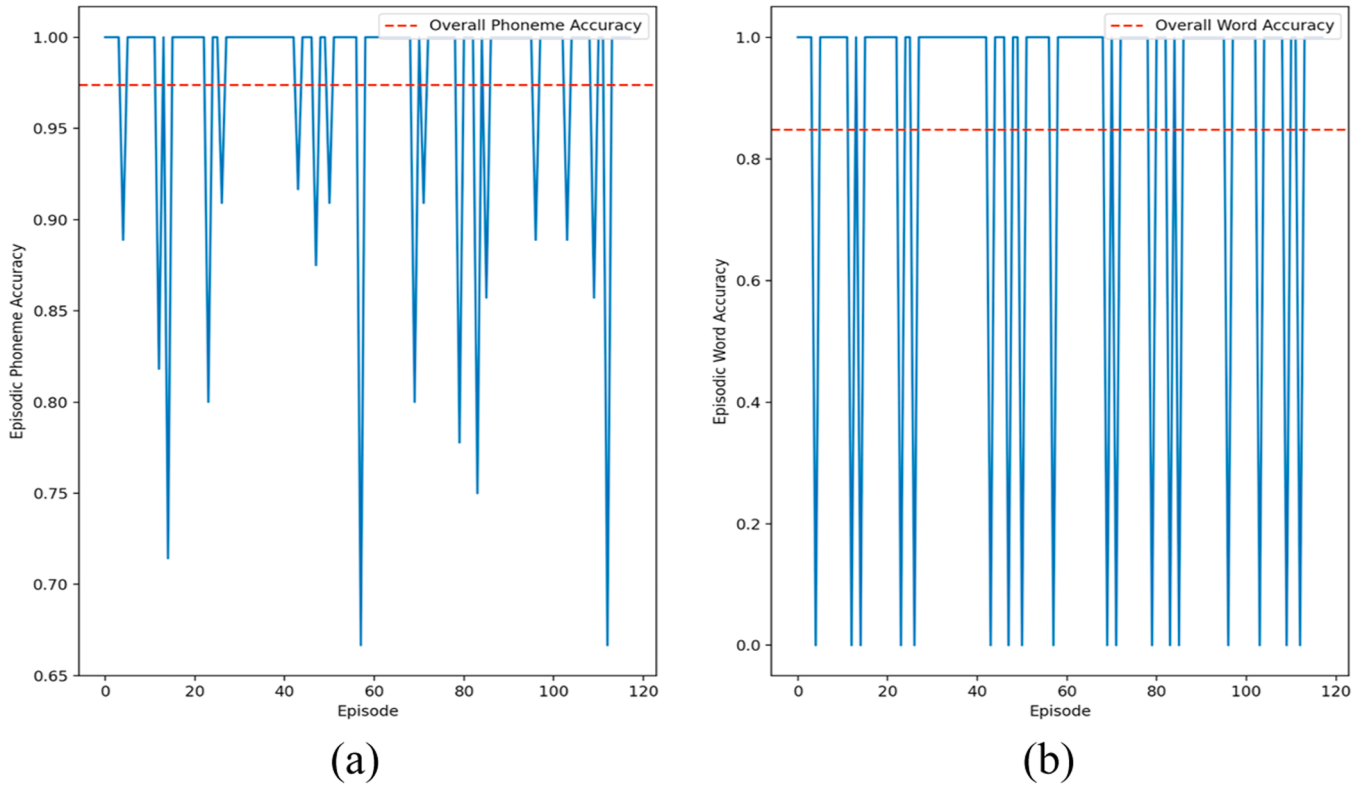


Fig. 5. (a) Episodic Phoneme-level Accuracy of Reinforcement Learning model, (b) Episodic Word-level Accuracy of Reinforcement Learning model.

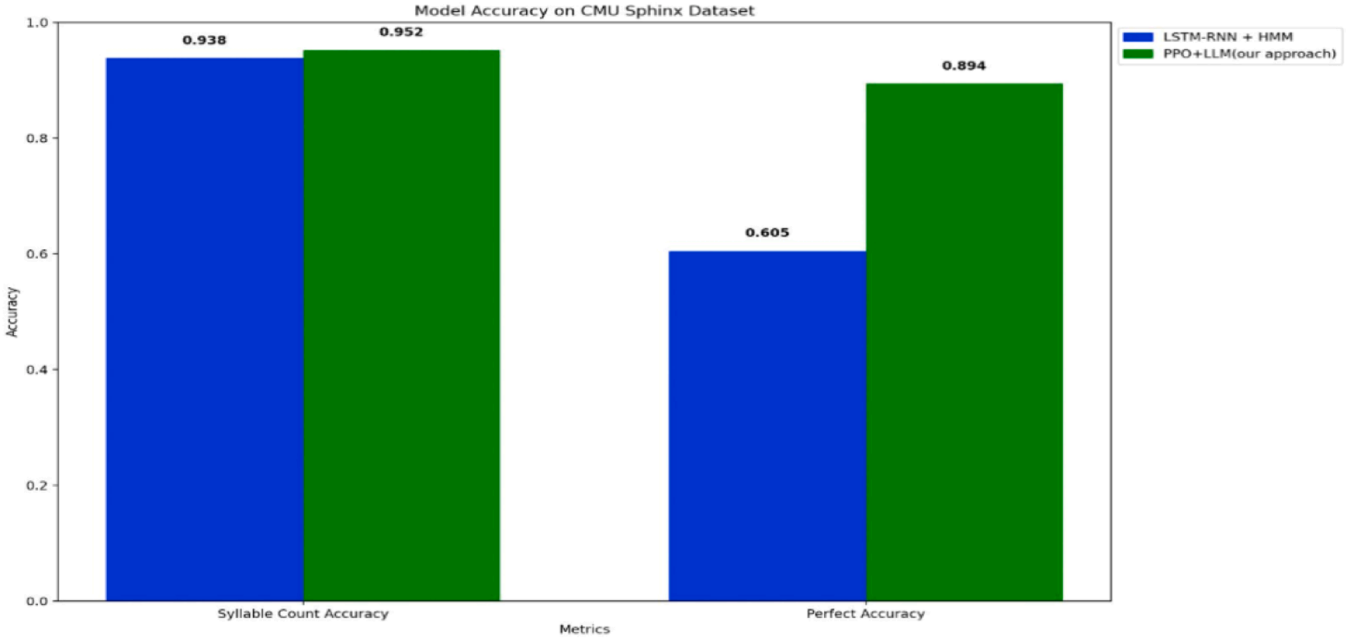


Fig. 6. Accuracy comparison of LSTM-RNN+HMM [7] and PPO + LLM model on CMU Sphinx dataset.

performance with other approaches, such as DNN on the TIMIT dataset, Speech-to-Speech (S2S) error generator with an encoder-decoder neural network (NN), on a combination of the TIMIT and LibriTTS datasets, Back-propagation(BP) and Radial Basis Function (RBF) based model on the RAVDESS dataset [23], Multi-Modal Large Language Model on the SpeechOcean762 dataset [24], and a DNN and Hidden Markov Model (HMM) based model [22]. Additionally, since the DNN + HMM model was trained on a custom spoken language dataset of English Language Learners, we evaluated its performance against our own dataset of random words, The results, depicted in Fig. 7, demonstrate our model's superior accuracy in detecting pronunciation errors using phonetic transcription.

As it can be observed, after training and testing on different datasets, our model's performance in detecting pronunciation errors using phonetic transcription is comparatively higher than the other given models. This is because our model is trained to use small steps to improve the phonetic transcription, thus reducing the scope for errors while testing the model with any user input. The same results have been mentioned in Table 5.

The performance variations across datasets can also be attributed to the specific characteristics of each dataset. For example, datasets like TIMIT, which are designed for phoneme-rich and structured speech, allow the system to achieve higher accuracy due to the clarity and consistency of pronunciation. In contrast, LibriTTS introduces sentence-

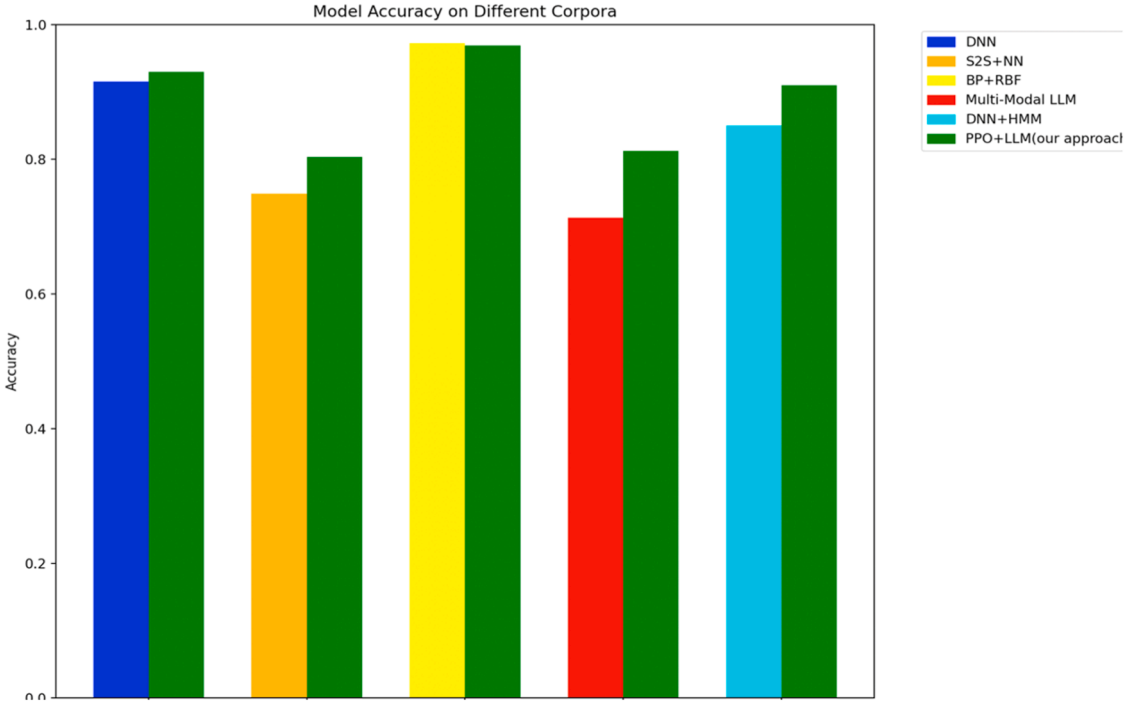


Fig. 7. Accuracy comparison of DNN[16], S2S + NN[17], BP + RBF[23], Multi-Modal LLM[24], DNN + HMM[22]and PPO + LLM model on respective datasets.

Table 5
Accuracy comparison of models for pronunciation error detection on different datasets.

Model	Dataset	Model's Accuracy	PPO+LLM's (Proposed System) accuracy
DNN[16]	TIMIT	0.915	0.93
S2S + NN [17]	TIMIT + LibriTTS	0.749	0.803
BP + RBF [23]	RAVDESS	0.972	0.969
Multi-Modal LLM [24]	SpeechOcean762	0.713	0.812
DNN + HMM [22]	Spoken samples from English Language Learners with diverse accents and backgrounds	0.85	0.91

Note that the accuracy of the model when applied on the TIMIT+LibriTTS dataset is lower than when applied on TIMIT dataset only. This is because the LibriTTS dataset consists of complete sentences and not individual words. This makes the model training harder and hence reduces the accuracy.

level complexity and varying linguistic structures, challenging the system's ability to generalize across longer inputs. Similarly, on emotional datasets like RAVDESS, the system's performance remains robust, with phoneme-level accuracy of 96.9 %. The reason for this is that the system, despite facing emotional speech patterns that introduce tonal variation, is designed to focus on phonetic transcription, which remains relatively stable across emotional states. The system's ability to adapt to these changes in pitch and tone allows it to achieve comparable results to other datasets. However, slight variations in accuracy may arise due to the emotional content and variability in speech delivery.

These results highlight the system's ability to handle diverse linguistic inputs effectively, but they also underscore areas for further optimization, such as improving performance in sentence-level and highly dynamic scenarios like emotional speech. Future work may focus on refining the model's adaptability to such variations, further enhancing the accuracy and robustness across all types of speech inputs.

To foster a positive and supportive learning environment, the Large Language Model module is designed to provide encouraging feedback and detailed explanations of pronunciation errors. We have employed sentiment analysis to assess the supportive nature of the Large Language Model's responses. Fig. 8 illustrates the sentiment trends in the feedback throughout the therapy sessions. The sentiment score usually ranges from −1 to +1, and our model's sentiment score never goes below 0. These results provide insights into the Large Language Model's capacity

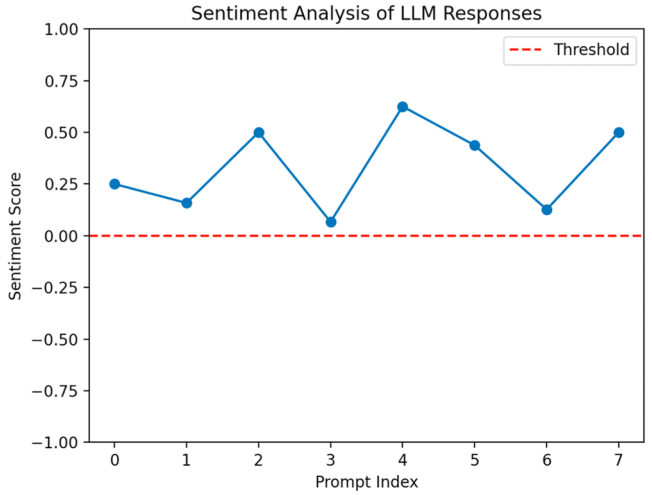


Fig. 8. Sentiment scores of Large Language Model's responses.

to maintain a positive and encouraging tone, even when the user does not perform the task well.

Advanced Automatic Speech Recognition systems, such as Google's Speech-to-Text API and others, have shown impressive results in transcription accuracy. However, these systems often struggle with the subtle nuances and dynamic speech patterns required for personalized pronunciation correction. In our system, the Large Language Model works in conjunction with the Automatic Speech Recognition module to ensure that the feedback provided is not only accurate but also contextually relevant to the user's specific pronunciation errors.

For instance, the combination of Reinforcement Learning and Large Language Model in our system allows for dynamic adaptation of therapy recommendations based on real-time user input. This iterative process ensures continuous improvement in pronunciation correction, something that static Automatic Speech Recognition systems cannot achieve.

4.3. Ablation study

In this subsection, we conducted experiments to evaluate the relative contributions of each module in the proposed PPO+LLM system—the Speech-to-Text module, the Reinforcement Learning module, and the Large Language Model module. Each module was systematically tested in isolation and in combination with others to quantify its impact on the overall system's performance. The Text-to-Speech module was not included in the ablation study, as it primarily changes the output format from text to audio, without affecting the accuracy or performance of the system. Therefore, the Text-to-Speech module does not influence the core functionality or the pronunciation correction process. As part of the ablation study the following configurations listed in Table 6 were tested and the accuracy is reported.

As observed in the Table 6, The PPO+LLM system, with all modules active, achieved the highest phoneme-level accuracy of 97.5 % and word-level accuracy of 84.7 %, demonstrating the complementary strengths of the Speech-to-Text, Reinforcement Learning, and Large Language Model modules working together. This configuration captures the user's pronunciation nuances, refines feedback dynamically, and provides context-aware corrections. When the Speech-to-Text module is removed, the system uses pre-transcribed phonemes (i.e., phonetic transcriptions of words without capturing the user's actual pronunciation errors). As a result, phoneme-level accuracy drops significantly to 72.5 %. Without capturing the user's pronunciation mistakes, the system cannot analyze how each user mispronounces words, leading to poor performance. The word-level accuracy also falls to 67.3 %, as the system is unable to refine the input dynamically based on individual user performance.

Removing the Reinforcement Learning module reduces the system's adaptability, leading to a drop in phoneme-level accuracy to 86.4 %. The RL module is essential for refining the feedback based on the user's ongoing performance, and without it, the system lacks the ability to provide dynamic, personalized corrections. Word-level accuracy drops to 73.2 % due to the absence of real-time refinement and adaptation. The user transcript directly generated by the Speech-to-Text module was

Table 6
Ablation study results: performance comparison of the full system and isolated modules.

System Configuration	Phoneme-Level Accuracy (%)	Word-Level Accuracy (%)
Full System (STT + RL + LLM)	97.5	84.7
Without STT (RL + LLM)	72.5	67.3
Without RL (STT + LLM)	86.4	73.2
Without LLM (STT + RL)	91.2	89.5
STT Only	89.0	76.9
RL Only	79.2	69.5
LLM Only	84.2	80.9

fed into the Large Language Model for feedback generation. This approach often failed to accurately identify pronunciation errors and produced misleading feedback. To overcome this, the Proximal Policy Optimization component was integrated and the system was able to refine its feedback generation process significantly. By leveraging it, the model incrementally learned to adapt its feedback based on user input, enabling more precise identification of pronunciation errors and personalized suggestions for improvement. For example:

- The system accurately detected nuanced errors in vowel sounds and consonant clusters.
- Feedback became contextually aligned with user pronunciation, avoiding false positives like those seen in the examples without Proximal Policy Optimization.

These shortcomings underscore the necessity of the Proximal Policy Optimization algorithm in refining feedback quality and ensuring accurate, adaptive guidance during pronunciation training.

The ablation study clearly shows that Allosaurus (Speech-to-Text module) is crucial for phonetic transcription, as its removal leads to a substantial decline in accuracy. The Reinforcement Learning module adds significant value by dynamically adapting to user-specific errors, improving the system's ability to provide personalized corrections. Finally, the Large Language Model module provides context-aware feedback, enhancing the system's performance in delivering nuanced corrections. While each module plays an important role, the integration of all three modules yields the highest performance, demonstrating the effectiveness of the proposed system.

4.4. Error analysis and model robustness

To evaluate the robustness of the proposed system, we performed a comprehensive error analysis using multiple datasets, including CMU Sphinx, TIMIT, LibriTTS, SpeechOcean762, and RAVDESS. This analysis focuses on understanding where the model misclassifies or struggles, particularly under different phonetic challenges, and evaluates its adaptability to various speech types and conditions. The analysis revealed specific areas where the model demonstrated high performance and robustness. The confusion matrix (Fig. 9) highlights areas of phoneme misclassification, particularly between phonetically similar sounds.

On the SpeechOcean762 dataset, which features non-native speakers, the model maintained competitive accuracy but exhibited a slightly higher misclassification rate for accents that deviated significantly from standard American English pronunciation. Despite this, its phoneme-level accuracy of 81.2 % surpassed that of the Multi-Modal Large Language Model, demonstrating its adaptability across diverse accents. Testing with the TIMIT dataset, augmented with background noise, revealed a slight drop in performance for the speech-to-text module. However, the overall system accuracy remained consistent, with a phoneme-level accuracy of 93.0 %, showcasing its resilience in challenging acoustic environments. When evaluated on emotional speech from the RAVDESS dataset, the system achieved a phoneme-level accuracy of 96.9 %. Despite tonal variations and emotional expressiveness affecting pronunciation, the model demonstrated robust performance by focusing on phonetic transcription, which remains relatively stable across emotional states.

As shown in Fig. 9, consonant clusters and multi-syllabic words exhibited slightly reduced accuracy. However, the system's phoneme-level accuracy consistently exceeded 90 % for most phonemes, validating its robustness across a variety of linguistic challenges.

These trends align with the dataset-specific findings and emphasize the robustness of the model while pinpointing specific areas for refinement, such as:

- Enhancing voicing distinction for voiced and voiceless phonemes.
- Improving the recognition of subtle articulatory differences in fricatives and stop consonants.

4.5. Efficiency analysis

4.5.1. Training time and memory usage

The training time and memory usage of the proposed system was evaluated under various configurations, measured for different numbers of episodes (50, 100, 200, and 500). While we were unable to find training time or memory usage data for the other models, the following results reflect the performance of our system under the same conditions.

As shown in Table 7, the training time increases with the number of episodes, and the memory usage increases with the number of episodes, reflecting the growing complexity of the system as it processes more data. The same have been clearly depicted in Fig. 10.

4.5.2. Response time comparison

To further evaluate the performance of the proposed system, the

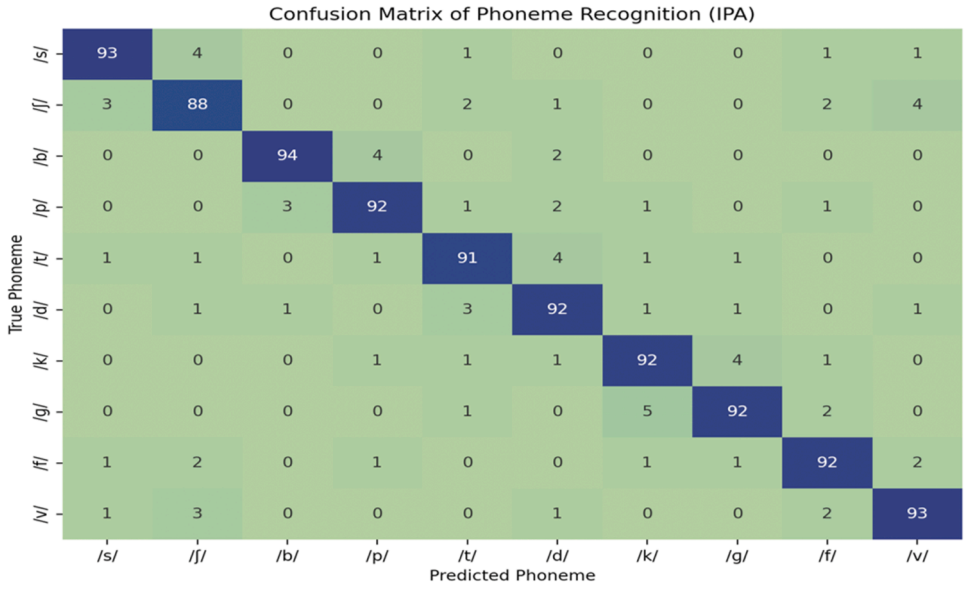


Fig. 9. Confusion Matrix for Phoneme-Level Performance (IPA Notation).

Table 7
Training time and memory usage for different numbers of episodes for our proposed model.

Number of Episodes	Training time (s)	Memory Usage (MB)
50	2.5	300
100	5.0	513
200	12.4	700
500	27.6	1032

response time was compared with the Back-propagation(BP) and Radial Basis Function (RBF) based model [23] and across different vocabulary sizes. This measures how quickly the system processes user input and provides feedback.

As it can be seen in Table 8, the response times for our PPO + LLM model are slightly higher than contemporary models, which is expected due to the added complexity of combining Reinforcement Learning and Large Language Model. However, these response times are still competitive and allow for efficient real-time interaction in pronunciation correction tasks. The same has been depicted in Fig. 11.

4.5.3. Computational challenges of reinforcement learning and large language model integration

The integration of Reinforcement Learning and the Large Language Model significantly impacts the computational demands of the system. The Reinforcement Learning component requires extensive training over multiple episodes to refine feedback based on user interactions, which can lead to increased training time. Each iteration involves processing large datasets to optimize the system's performance in real-time, which is a time-consuming process. While the Large Language Model is used to generate context-aware feedback, its role is primarily during the runtime to process the user input and generate detailed feedback. The memory usage captured reflects the overall system's requirements during both training and inference phases, which include not only Reinforcement Learning but also the Large Language Model and Speech-to-Text modules. The combination of these three modules requires substantial memory to process and store intermediate data during training and real-time user interaction.

Thus, the overall memory consumption is a sum of the memory usage across all modules and reflects the integrated system's demands. The Large Language Model's contribution is significant in generating accurate and personalized feedback, but the memory usage does not store Large Language Model results separately, as it is part of the complete system processing.

While the proposed system demonstrates significant advancements in personalized pronunciation correction, a few limitations and challenges do exist such as a heavy dependency on input data quality, latency issues during real time processing and scalability to large volume of data. Despite these challenges, the system's foundational focus on personalized feedback, iterative learning, and user-centric design ensures its overall effectiveness. Ongoing improvements to the Large Language Model module will further enhance its reliability and performance.

5. Conclusion and future work

This work presents a novel, automated system for personalized pronunciation assessment and improvement, integrating advanced technologies such as Proximal Policy Optimization and Large Language Models. The Proximal Policy Optimization algorithm effectively evaluates pronunciations, while the Large Language Model provides nuanced feedback and personalized explanations, fostering trust and engagement through positive reinforcement, as demonstrated by sentiment analysis results. This combination of Reinforcement Learning, Large Language Models and natural language processing holds significant promise for advancing pronunciation correction tools and personalized language learning.Beyond its technical contributions, the system's broader implications highlight its potential to transform the landscape of speech therapy and language education. By offering dynamic, real-time feedback tailored to individual learners, the system provides an accessible alternative to traditional therapy methods, making high-quality pronunciation correction available to a wider audience. Additionally, the modular architecture ensures adaptability across different languages and dialects, addressing the global demand for scalable and culturally inclusive learning tools.

However, while the system demonstrates strong performance, some challenges remain that could impact its scalability and real-world deployment. The dependency on high-quality input data underscores

Table 8
Response time comparison between our system and BP+RBF model (in seconds).

Vocabulary Size	BP + RBF[23]	PPO+LLM (Proposed System)
Group 1 (897)	0.347	0.37
Group 2 (798)	0.259	0.30
Group 3 (856)	0.323	0.35
Group 4 (791)	0.251	0.28
Group 5 (912)	0.355	0.40

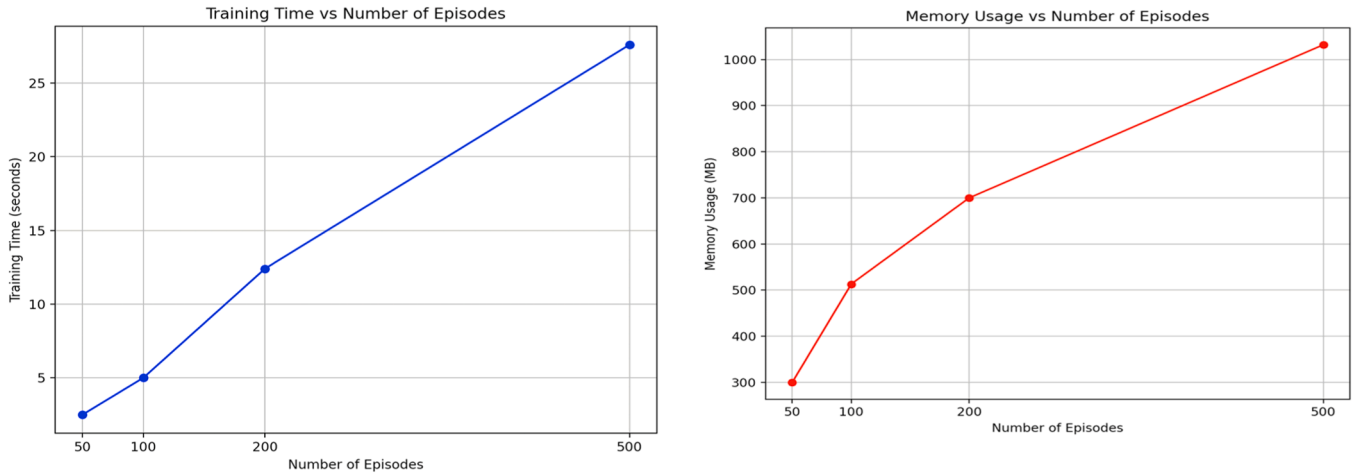


Fig. 10. (a) Training time (in seconds) of Reinforcement Learning model, (b) Memory usage (in MB) of our proposed model.

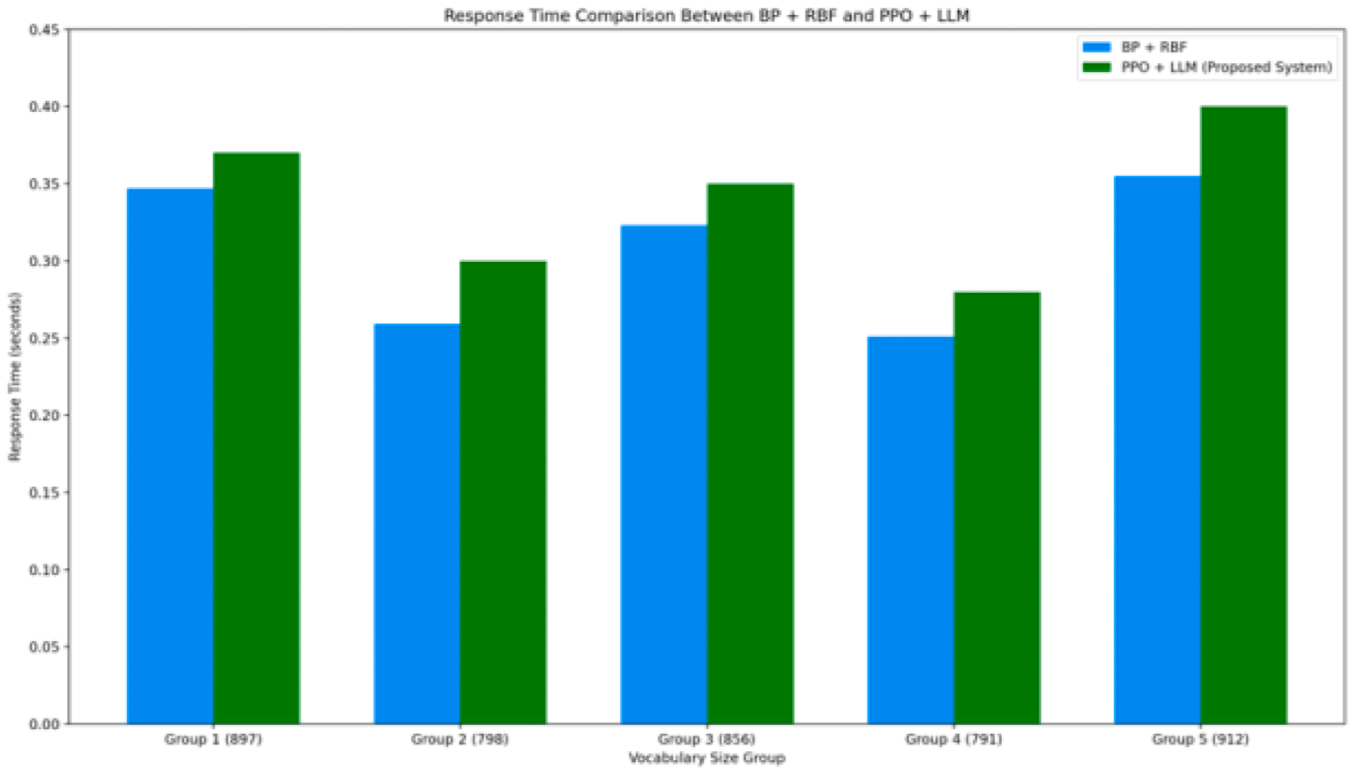


Fig. 11. Response time comparison between BP + RBF[23]and PPO + LLM (proposed system).

the need for robust noise-cancellation techniques, particularly in noisy or resource-constrained environments. Furthermore, the system's reliance on language-specific datasets requires ongoing effort to expand its multilingual capabilities. Latency issues and computational demands also present obstacles, especially in live, large-scale applications. Future work will be focussed towards addressing these limitations and ensuring the system's effectiveness across diverse linguistic and cultural contexts.

Declarations

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to A. S.

CRediT authorship contribution statement

Ritika Lakshminarayanan: Writing – review & editing, Writing – original draft, Validation, Conceptualization. **Ayesha Shaik:** Writing – review & editing, Writing – original draft, Validation, Supervision, Formal analysis, Conceptualization. **Ananthakrishnan Balasundaram:** Writing – review & editing, Writing – original draft, Validation, Methodology, Investigation, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

The authors would like to thank Centre for Cyber Physical Systems and School of Computer Science and Engineering, Vellore Institute of Technology, Chennai for giving the support and encouragement to proceed with the research and produce fruitful results.

Data availability

The data will be provided by the corresponding author on request.

References

[1] J.C. Shanbal, M.S. Reddy, Distribution of communication disorders in primary school children, J. All India Inst. Speech Hear. (2015) 34.
[2] National Institute on Deafness and Other Communication Disorders (NIDCD). (2016, May 19). Quick Statistics About voice, speech, Language. Retrieved from https://www.nidcd.nih.gov/health/statistics/quick-statistics-voice-speech-language.
[3] M.H. Franciscatto, M.D. Del Fabro, J.C.D. Lima, C. Trois, A. Moro, V. Maran, M. Keske-Soares, Towards a speech therapy support system based on phonological processes early detection, Comput. Speech Lang. 65 (2021) 101130.
[4] Y. Yuan, X. Liu, An empirical study of the effect of ASR-supported English reading aloud practices on pronunciation accuracy, in: Proceedings of the Technology in Education. Innovations for Online Teaching and Learning: 5th International Conference, ICTE 2020, Macau, China 5, 2020, pp. 75–87. August 19-22, 2020Revised Selected PapersSpringer Singapore.
[5] M. Malik, M.K. Malik, K. Mehmood, I. Makhdoom, Automatic speech recognition: a survey, Multimed. Tools Appl. 80 (2021) 9411–9457.
[6] K.F. Lee, H.W. Hon, R. Reddy, An overview of the SPHINX speech recognition system, IEEe Trans. Acoust. 38 (1) (1990) 35–45.
[7] P. Nandal, Y. Kadian, S. Upadhyay, B.P. Mudgal, Pronunciation accuracy calculator using machine learning, in: Proceedings of the 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), 2021, pp. 1128–1133. IEEE.
[8] J. Fu, Y. Chiba, T. Nose, A. Ito, Automatic assessment of English proficiency for Japanese learners without reference sentences based on deep neural network acoustic models, Speech Commun. 116 (2020) 86–97.

[9] Z. Zhang, Y. Wang, J. Yang, Text-conditioned transformer for automatic pronunciation error detection, Speech Commun. 130 (2021) 55–63.
[10] V.C.W. Cheng, V.K.T. Lau, R.W.K. Lam, T.J. shan, P.K. Chan, Improving English phoneme pronunciation with automatic speech recognition using voice chatbot, in: Proceedings of the Technology in Education. Innovations for Online Teaching and Learning: 5th International Conference, ICTE 2020, Macau, China, August 19-22, 2020, Revised Selected Papers 5, 2020, pp. 88–99. Springer Singapore.
[11] B. Min, H. Ross, E. Sulem, A.P.B. Veyseh, T.H. Nguyen, O. Sains, D. Roth, Recent advances in natural language processing via large pre-trained language models: a survey, ACM Comput. Surv. 56 (2) (2023) 1–40.
[12] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, R. Lowe, Training language models to follow instructions with human feedback, Adv. Neural Inf. Process. Syst. 35 (2022) 27730–27744.
[13] B. Lin, Reinforcement Learning and bandits for speech and language processing: tutorial, review and outlook, Expert Syst. Appl. (2023) 122254.
[14] H. Chung, H.B. Jeon, J.G. Park, Semi-supervised training for sequence-to-sequence speech recognition using Reinforcement Learning, in: Proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN), 2020, pp. 1–6. IEEE.
[15] s. Min, J. Wang, Exploring the integration of large language models into automatic speech recognition systems: an empirical study, in: Proceedings of the International Conference on Neural Information Processing, 2023, pp. 69–84. Singapore: Springer Nature Singapore.
[16] M. Shahin, B. Ahmed, Anomaly detection based pronunciation verification approach using speech attribute features, Speech Commun. 111 (2019) 29–43.
[17] D. Korsekwa, J. Lorenso-Trueba, T. Drugman, B. Kostek, Computer-assisted pronunciation training—speech synthesis is almost all you need, Speech Commun. 142 (2022) 22–33.
[18] O. Gillath, T. Ai, M.S. Branicky, S. Keshmiri, R.B. Davison, R. Spaulding, Attachment and trust in artificial intelligence, Comput. Hum. Behav. 115 (2021) 106607.
[19] F. Gille, A. Jobin, M. Ienca, What we talk about when we talk about trust: theory of trust for AI in healthcare, Intell. Based. Med. 1 (2020) 100001.
[20] X. Li, S. Dalmia, J. Li, M. Lee, P. Littell, J. Yao, F. Metse, Universal phone recognition with a multilingual allophone system, in: Proceedings of the ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 8249–8253. IEEE.
[21] P. Schmidt, F. Biessmann, T. Teubner, Transparency and trust in artificial intelligence systems, J. Decis. Syst. 29 (4) (2020) 260–278.
[22] M.S. Lara, R. Subhashini, C. Shiny, J.C. Lawrance, S. Prema, S. Muthuperumal, Constructing an AI-assisted pronunciation correction tool using speech recognition and phonetic analysis for ELL, in: Proceedings of the 2024 10th International Conference on Communication and Signal Processing (ICCSP), 2024, pp. 1021–1026. IEEE.
[23] X. Wang, Research on oral english learning system integrating AI speech data recognition and speech quality evaluation algorithm, J. Electr. Syst. 20 (5s) (2024) 2466–2477.
[24] Fu, K., Peng, L., Yang, N., & Zhou, S. (2024). Pronunciation assessment with multi-modal large language models. arXiv preprint arXiv:2407.09209.
[25] Google Cloud. "Speech-to-Text." Available at: https://cloud.google.com/speech-to-text.
[26] Hannun, A., et al. "Deep Speech: scaling up end-to-end speech recognition." arXiv preprint arXiv:1412.5567 (2014).
[27] Radford, A., et al. (2022). Robust speech recognition via large-scale weakly supervised pre-training. OpenAI. https://openai.com/research/whisper.