

Enhancing Speech Clarity in Individuals with Articulation and Phonological Disorders using Advanced Computational Model

Ezgi Demir¹, S. Rubin Bose^{2,*}, Mouhamadou Fall³, Cao Siyu⁴, Li Chengru⁵, Ankan Basak⁶

¹Department of Artificial Intelligence, Lancaster University, Lancaster, Lancashire, England, United Kingdom.

²Department of Computer Science and Engineering, SRM Institute of Science and Technology, Ramapuram, Chennai, Tamil Nadu, India.

³Department of Intelligent Transportation Systems, LITIS - IT Laboratory, Information Processing and Systems, University of Rouen Normandy, Mont-Saint-Aignan, Normandy, France.

⁴Department of Computing and Information Systems, Singapore Management University, Basah, Singapore.

⁵Department of Information and Communication Engineering, The University of Tokyo, Bunkyo-ku, Tokyo, Japan.

⁶Department of Computational Science and Engineering, McMaster University, Hamilton, Ontario, Canada.

ezgi.demir@nisantasi.edu.tr¹, rubinbos@srmist.edu.in², mouhamadou.fall@insa-rouen.fr³, siyu.cao.2026@phdcs.smu.edu.sg⁴, lichengru@logos.t.u-tokyo.ac.jp⁵, basaka3@mcmaster.ca⁶

Abstract: The paper describes a pronunciation feedback and exercise-generation tool designed to enhance language acquisition through real-time phoneme analysis. The system evaluates the correctness of a student's pronunciation using cutting-edge speech recognition and Generative AI (GenAI) technologies, then creates customised exercises to address particular phonetic issues. The method combines a simple UI that lets students engage with the system and get immediate speech feedback. The system identifies differences between the target and spoken pronunciation by analysing the phonemic structure of the learner's utterances and provides insights into error patterns. These ideas then inform tailored exercises designed to strengthen troublesome phonemes. GenAI in the system enables adaptive learning, in which exercise difficulty and emphasis change based on the student's development. This adaptive strategy ensures the system can support students at different skill levels. The findings suggest that consistent, data-driven practice can significantly improve pronunciation, enabling more efficient and accessible language acquisition. This paper examines the system's design, implementation, and potential impact on contemporary language teaching.

Keywords: Speech Therapy; Speech-Language Pathology (SLP); Speech Disorder; Adaptive Strategy Guarantees; Data-Driven Practice; Contemporary Language Teaching.

Received on: 12/07/2025, **Revised on:** 23/09/2025, **Accepted on:** 20/10/2025, **Published on:** 03/06/2026

Journal Homepage: <https://www.fmdbpublish.com/user/journals/details/FTSBE>

DOI: <https://doi.org/10.69888/FTSBE.2026.000789>

Cite as: E. Demir, S. R. Bose, M. Fall, C. Siyu, L. Chengru, and A. Basak, "Enhancing Speech Clarity in Individuals with Articulation and Phonological Disorders using Advanced Computational Model," *FMDB Transactions on Sustainable Biomedical Engineering*, vol. 1, no. 2, pp. 73–85, 2026.

Copyright © 2026 E. Demir *et al.*, licensed to Fernando Martins De Bulhão (FMDB) Publishing Company. This is an open access article distributed under [CC BY-NC-SA 4.0](https://creativecommons.org/licenses/by-nc-sa/4.0/), which permits copying, remixing, redistributing, transformation, and building upon the material in any medium so long as the original work is properly cited.

1. Introduction

Speech and language disorders encompass a broad range of conditions that affect millions of people of all ages, compromising vital aspects of human interaction, including communication, social integration, academic success, and overall quality of life

*Corresponding author.

[1]. Among the many ways to address these issues, speech-language therapy has consistently shown effectiveness in identifying, treating, and sometimes reversing the functional constraints linked to these conditions [2]. However, especially in under-resourced and rural areas, factors such as a small number of trained speech-language pathologists (SLPs), geographic isolation, infrastructure deficiencies, and financial limitations seriously limit speech therapy's global accessibility and scalability, even when it is clinically effective [3]. The difference between the growing worldwide need for SLP services and the supply of professionals calls for the investigation of scalable, affordable, and technically strong substitutes that can augment and improve conventional therapy models. Recent developments in artificial intelligence (AI), particularly in natural language processing (NLP), machine learning (ML), and automatic speech recognition (ASR), offer transformative possibilities for delivering speech therapy [4]. These technologies enable smart, flexible systems that can deliver real-time, tailored, interactive therapy sessions beyond the limits of clinical settings. By lowering barriers to entry, providing continuous phoneme-level feedback, automating routine assessments, and generating personalized therapy materials based on the user's speech patterns and therapeutic goals, AI-powered systems can broaden access to therapy. By revealing the fragility of traditional, face-to-face therapeutic models, which lockdowns and social distancing policies drastically disrupted, the COVID-19 epidemic further highlighted the need for remote healthcare solutions. These disruptions underscored the need for resilient, telehealth-compatible solutions and caused therapy regression, particularly among paediatric and geriatric populations [5].

Although the larger telehealth market has expanded quickly, AI-based speech therapy systems integrated into clinical workflows remain scarce, and few are sufficiently granular and adaptive. Current tools often lack phoneme-level error detection, context-aware feedback generation, and adaptive content delivery [6]. Many also require significant manual SLP input, limiting scalability and undermining their value as autonomous or semi-autonomous support systems. This study offers a technically sophisticated, AI-powered virtual speech therapy support system meant to integrate smoothly into current clinical and educational systems and give patients a dynamic, data-driven therapeutic experience in response to these deficiencies. A high-fidelity phoneme recognition engine, trained on several linguistic datasets to ensure accurate and robust articulation analysis across many dialects and age groups, is central to the system's architecture [7]. This engine lets users identify, understand, and correct misarticulations during self-guided practice sessions through real-time auditory and visual feedback. A dynamic content generation module also customises practice materials based on ongoing user progress and specific speech issues, enabling the therapy process to adapt in synchrony with the person's developmental path [8]. Lightweight, computationally efficient, and compatible with standard audio recording formats, the proposed system is designed for inclusive, practical deployment, making it feasible for use in a variety of environments—from urban clinics and school-based intervention programs to remote, resource-constrained home settings [9]. Its mobile and web use requires only a basic internet-enabled device, promoting easy access. For SLPs, the system is a clinical augmentation tool that supports routine diagnostic evaluations, longitudinal progress tracking, and automated report generation, maximising practitioner time and allowing greater caseload efficiency without sacrificing therapeutic quality [10].

Designed with ethical data stewardship in mind, strong encryption and anonymisation techniques ensure that sensitive speech data used for feedback and progress tracking remains secure and compliant with applicable data protection laws [11]. Furthermore, the platform's aggregated, anonymised data could serve as a useful corpus for cross-sectional and longitudinal studies on speech-language development and therapy effectiveness. The system aims to improve the accuracy and repeatability of therapeutic interventions through automation, reducing human error and maximising consistency [12]. From a systems integration perspective, the development uses a modular, interoperable software architecture that allows future expansion into multilingual support, integration with electronic health records (EHR), and compatibility with other educational or assistive technologies [13]. The platform stays lightweight enough for mobile deployment while maintaining high accuracy and responsiveness by applying deep learning models optimised for constrained environments and fine-tuned using domain-specific corpora. Importantly, the platform does not seek to replace the knowledge of licensed speech-language pathologists but to function as an intelligent co-therapist, therefore improving the scope, frequency, and precision of therapy and enabling SLPs to concentrate on higher-order clinical decision-making and patient involvement [14]. Overall, this study offers a thorough, AI-enhanced virtual therapy tool that addresses the multidimensional issues of accessibility, individualisation, and scalability in speech-language therapy. Aiming to close the gap between clinical effectiveness and real-world applicability, it reflects a confluence of technological innovation and therapeutic best practices. While establishing a standard for future integration of artificial intelligence into specialised therapeutic areas, the main objective is to improve communication and quality of life for people with speech and language problems, particularly those marginalised by current healthcare systems [15].

2. Materials and Methods

The proposed system, an AI-driven platform, assesses and improves pronunciation skills through customised remedial exercises and automated feedback. It comprises several connected parts, each responsible for specific tasks, as shown in Figure 1. Users access the system through a web-based interface built using Flask and ngrok, deployed for secure tunnelling. After logging in, users see daily sentence exercises from a selected database. Curated to fit the user's skill level, these sentences progressively get more complicated to support progressive learning. Designed to enable smooth interaction, the user interface is simple. It

uses the browser's Media Recorder API to record user speech during pronunciation practice. Once recorded, the audio is uploaded to Google Drive for later processing. To preserve data integrity during transfer and storage, the recordings are base64 WAV encoded. The audio processing pipeline starts by pulling these recorded files from cloud storage. The system performs necessary preprocessing using libraries including Librosa and Pydub. These include amplitude normalisation to equalise volume across recordings, silence trimming to remove pauses, and denoising to reduce background noise. This guarantees consistency in the audio data, a crucial step for dependable feature extraction and analysis. After preprocessing, the system extracts key acoustic characteristics. Using tools such as torchaudio and TensorFlow, it generates Log-Mel spectrograms and Mel-Frequency Cepstral Coefficients (MFCCs). These features capture the speech signal's spectral properties and temporal dynamics. The combined feature matrix serves as the basis for pronunciation analysis. Simultaneously, the system analyses the text material of the pronunciation drills. To ensure consistent formatting, the system normalises the reference text by removing special characters, adjusting case sensitivity, and tokenising it into analysable units. Researchers then use tools such as the CMU Pronouncing Dictionary (CMUdict) and Phonetisaurus to apply Grapheme-to-Phoneme (G2P) conversion.

The phonemic ground truth for comparison generated by this conversion is the corresponding phonemic sequence of the cleaned text. The Sequence-to-Sequence (Seq2Seq) model, built in PyTorch, is central to the system. This model matches the phonemic sequences with the acoustic characteristics. It uses a Long Short-Term Memory (LSTM) encoder to process the acoustic feature matrix in both forward and backward directions. Also, the LSTM decoder is enhanced with attention mechanisms that let the model, during prediction, selectively focus on relevant input segments. Essential for precise phoneme prediction, this architectural design lets the model capture context and long-range dependencies. The decoder generates a forecasted phoneme sequence, and the phonemic ground truth is then used to evaluate the accuracy of the user's pronunciation. By matching the predicted phoneme sequence with the target sequence, the system conducts phoneme-level deviation analysis. It measures deviation using dynamic time warping (DTW), edit distance, and Phoneme Error Rate (PER). While edit distance offers a simple tally of insertion, deletion, and substitution errors, DTW aligns sequences of different lengths. PER provides a standardised error measure to evaluate pronunciation performance across sessions. A user-specific phoneme error log tracks past performance and flags consistent mistakes, thereby storing the study results. Data visualisation tools like Matplotlib, Plotly, and Seaborn then visualise these findings. The graphical reports show trends over time and highlight specific phonemes that need focus, enabling teachers and students to track progress. The system combines a Generative AI component driven by the GEMINI API to offer remedial teaching. This module examines the error log to find ongoing pronunciation issues. It then creates tailored remedial exercises meant to address the user's weak areas. Designed to support proper pronunciation patterns, these exercises are linguistically varied and contextually pertinent.

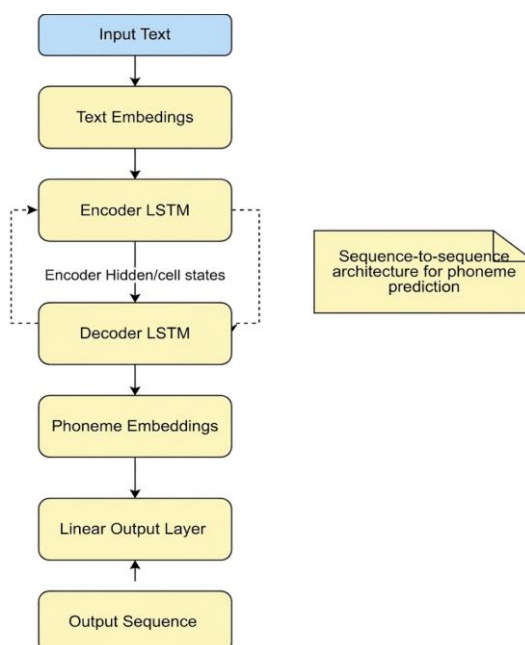


Figure 1: LSTM model architecture

In addition to being stored in a separate database, the tailored exercises are also planned for inclusion in the user's next practice sessions. Continuity and reinforcement are fundamental to language acquisition; this process supports both. Produced material encourages gradual improvement through incremental difficulty changes, so it adapts to the user's evolving skill level. Figure 1 visually depicts the modular pipeline used by the pronunciation feedback and personalisation engine. The system's core

components are the User Interface (UI), audio and text processing systems, feature extraction tools, the Seq2Seq pronunciation model, and the Generative AI module. The system's modularity makes it suitable for clinical settings, educational institutions, and language training centres. The system's scalability and ability to integrate new features or external APIs smoothly support this. Figure 1 shows a top-level depiction of a sequence-to-sequence (Seq2Seq) model for phoneme prediction. Figure 1 begins with "Input Text" and proceeds through several steps to produce an "Output Sequence." Every box in Figure 1 illustrates a different step within the processing chain. The input sequence is first translated into "Text Embeddings," then fed into an "Encoder LSTM." The encoder reads the input sequence and produces hidden and cell states, which carry context across the sequence. Figure 2 shows the full workflow of the pronunciation assessment and individualised exercise generation system, designed to give users real-time, focused feedback as they work to improve their pronunciation. The workflow begins when the user receives a daily sentence exercise. The system selects this activity to match the user's current knowledge level and learning needs. The user completes the activity by reading the sentence aloud. The system records the spoken input and automatically sends it to cloud storage.

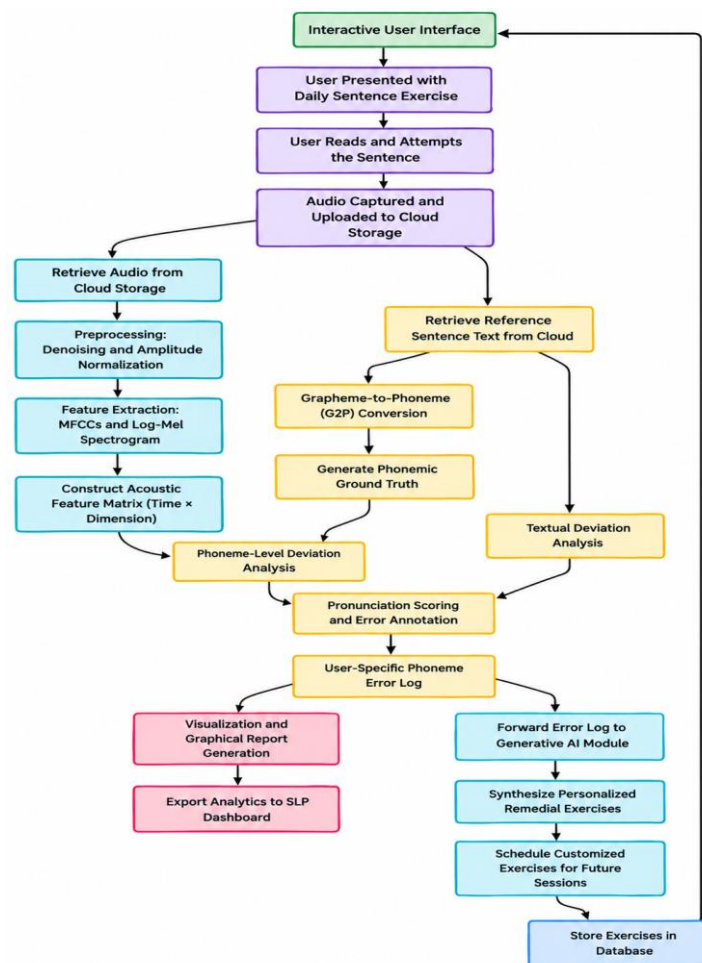


Figure 2: System architecture for pronunciation

The system retrieves the audio and the corresponding reference text from the cloud, then starts the dual-track preprocessing sequence. Meanwhile, the text is cleaned and converted into its phonemic representation using G2P conversion. Denoising and normalising follow the audio. The system employs the Seq2Seq model to match the extracted acoustic characteristics with the phonemic ground truth. This follows individual processing of the acoustic and textual data. Generating phoneme predictions based on the acoustic input, the attention-enabled decoder then compares them to the reference phonemes. Deviation analysis helps to find mispronounced phonemes and calculate a total pronunciation score. Storing this study in the phoneme error log allows users to build a historical record of their performance. The system generates feedback reports with easy-to-understand graphs and heatmaps. These reports graphically summarise errors and areas for improvement. These reports are designed to be readily available to both students and teachers, improving interpretability and offering actionable insights. The generative artificial intelligence module then gets the noted pronunciation mistakes. This component uses the error data to create tailored remedial exercises focused on the phonemes causing concern. Scheduled dynamically, these exercises remain part of the user's

practice until the user reaches a satisfactory level of mastery. The system is also updating a dashboard for language teachers and Speech-Language Pathologists (SLPs). This dashboard shows thorough analysis of the user's development.

These statistics include historical trends, phoneme-level breakdowns, and overall progress measures. Clinicians and teachers can use this information to customise treatments, plan reviews, and provide more targeted support. The system was designed with user experience and effective teaching approaches in mind. While longitudinal tracking supports long-term growth, real-time feedback enables quick correction. By combining AI-driven analytics with user-centred design, the system offers a new approach to speech therapy and language acquisition. This combined pipeline has driven notable progress in automated language teaching. User-friendly interfaces and tailored material delivery complement machine learning models at the forefront of technology. The modular design, which ensures adaptability across use cases, can support second language acquisition, accent reduction, and speech disorder intervention, among other applications. From user input through feedback generation and remedial planning, Figure 2 shows the complete end-to-end system architecture and highlights each phase. This methodical flow enables seamless data management and ensures reproducible results, which is crucial for educational and clinical validation. Ultimately, the AI-driven pronunciation evaluation and feedback tool presented in this paper offers language students a strong, scalable, and individualised solution.

By combining generative artificial intelligence, machine learning, and acoustic signal processing, the aim is to offer a thorough educational experience. Its modular design and real-time analytics make it well-suited for modern uses in language education and speech therapy. Future improvements could include integrating self-supervised learning models like wav2vec 2.0 or HuBERT and adding multilingual support for broader use. In addition, later versions of the system intend to incorporate adaptive learning approaches powered by reinforcement learning. This would help the platform optimize its strategies to improve teaching and learning outcomes by adapting to user interactions. For example, if a learner consistently struggles with certain phonemes and prosodic features like intonation or stress, the system may adjust exercise scheduling or change difficulty levels in real time, promoting more efficient and engaging learning experiences. To enhance equity and inclusivity, researchers are also exploring ways to accommodate multiple regional accents and dialects. This entails training the Seq2Seq model on broader speech datasets representing various linguistic backgrounds, thereby improving the system's generalisability. Moreover, providing voiced synthesis materials that demonstrate correct pronunciation helps users, particularly learners, analyse sounds accurately. Combined with highlighting of relevant symbols, this approach can make the exercise more effective and help ingrain the right speech patterns. Security and privacy remain key priorities. All user data, including audio recordings and progress logs, are securely stored and encrypted in accordance with international data protection regulations like GDPR. Future versions.

3. Experimental Setup

3.1. Training

The training approach focuses on fine-tuning a pre-trained sequence-to-sequence (Seq2Seq) model based on Long Short-Term Memory (LSTM) networks for speech-to-phoneme conversion to meet the needs of a virtual speech therapy system. The main goal is to generate very accurate phoneme sequences from orthographic input—i.e., standard written text—which can then be applied in a downstream pronunciation feedback system. The model uses transfer learning rather than starting from scratch, which is time-consuming and computationally costly. Domain-specific data from the speech therapy context adapt a pre-trained Seq2Seq LSTM network, improving generalisation and robustness to user-specific pronunciation challenges. The selected model design runs at a character-to-phoneme granularity. Specifically, the output is a sequence of phonemes, and the input sequence is a string of characters representing the word. This fine-grained method ensures high fidelity in handling unusual orthographies common in English. Comprising two primary parts—an encoder and a decoder—the model uses two stacked LSTM layers, each with a hidden size of 128 units. The character-level input passes through a trainable embedding layer that maps characters to dense vector representations before reaching the encoder. This embedding layer helps the model to better capture the subtleties of English spelling-to-sound rules by allowing it to learn relationships and contextual use of characters. The decoder's internal state is then initialised with the encoded context vectors, comprising the encoder's last hidden and cell states. During training, the decoder's own trainable embedding layer converts phoneme inputs into dense vectors. A softmax layer predicts the next phoneme in the sequence at each decoding step. During training, teacher forcing balances model direction and autonomous learning.

Specifically, with a 0.5 chance, the decoder receives either the ground-truth phoneme from the training dataset or the phoneme the model predicted at the prior timestep. Using a stochastic teacher-forcing schedule, the model can learn from mistakes without over-relying on the ground-truth sequence, improving robustness during inference. Padding tokens are ignored in the loss computation to guarantee computational efficiency and efficient gradient propagation. Applied over all non-padding sites in the predicted phoneme sequence, the loss function used is the conventional cross-entropy loss. The Adam optimiser uses a learning rate of 0.001, leveraging its adaptive learning-rate characteristics to speed convergence. Over 10 epochs, the training

process uses GPU acceleration to reduce training time and enable larger batch sizes. The Carnegie Mellon University Pronouncing Dictionary (CMUdict), a commonly used lexicon in speech and language processing, provides the dataset used for training and assessment. The dataset is preprocessed by standardising word formats, deleting duplicates, and normalising phoneme sequences to preserve uniformity. A 90:10 split divides the data into training and testing subsets. This separation ensures the model is validated on unseen examples, simulating real-world deployment. Training uses a batch size of 64, empirically chosen to balance memory use and convergence speed. Researchers track several measures throughout training to monitor learning progress and generalisation. Chief among these is the Phoneme Error Rate (PER), a domain-specific analogue to conventional edit distance measures. PER is the normalised Levenshtein distance between the predicted and ground-truth phoneme sequences. This covers insertions, deletions, and substitutions, and strongly indicates how precisely the model captures phonological data.

3.2. Evaluation

The performance of the fine-tuned Seq2Seq model is assessed exhaustively using both quantitative and qualitative methods. The Phoneme Error Rate (PER), which directly measures the model's capacity to generate correct phoneme sequences given text input, remains the main quantitative measure. PER is determined by totalling the edits needed to transform the predicted sequence into the reference sequence, normalised by the length of the reference. Total accuracy is a more intuitive measure, calculated by subtracting the resulting value from 1 and multiplying by 100. For example, a PER of 0.12 indicates overall phoneme-sequence accuracy of 88%. Quantitative testing is done only on the held-out 10% of the CMUdict dataset, which the model did not see during training. This is critical in real-world speech therapy applications where users may speak a varied vocabulary, including out-of-vocabulary (OOV) terms; it ensures the model is not overfitting and can generalise effectively to new words. Beyond PER, the system produces loss curves and accuracy graphs every epoch to provide insight into model convergence. These curves indicate possible overfitting, underfitting, or training instability. Enabled attention weights—when using attention mechanisms in a variant of the architecture—are also visualised to show which character-level inputs are most influential in deciding each phoneme output. This increases explainability and helps to correct misclassifications. Qualitative assessment uses manual inspection of the model's phoneme forecasts for a chosen set of representative terms, including "chatgpt," "language," and "therapy." Researchers selected these words for their phonological complexity and relevance to the application domain. Fidelity is assessed by comparing every forecasted phoneme sequence with a validated ground-truth sequence.

Misalignments are examined to find systematic errors—for example, consistent misunderstanding between similar-sounding phonemes like /s/ and /ʒ/. Researchers use user-supplied speech recordings in real-world testing to further evaluate the model's practicality. The Google Speech Recognition API, which transforms speech to text in noisy or casual speech settings, transcribes these audio samples. The transcriptions produced are sent to the model to create phoneme sequences, therefore simulating use in a virtual speech therapy session. This end-to-end test assesses how well the model performs under less-than-ideal input conditions, such as slurred speech, mispronunciations, and background noise—factors usually encountered in online therapeutic settings. Empirical PER in these circumstances is then calculated by comparing the model's phoneme outputs under real-world speech scenarios against the ideal phoneme representations. Furthermore, a phoneme class-wise study identifies classes where the model underperforms, such as diphthongs or rhotic vowels, which are frequent trouble spots for speech therapy participants. Insights from this study shape future model iterations and support curriculum development for tailored remedial activities. Overall, the fine-tuned Seq2Seq LSTM model consistently transforms character-level text inputs into correct phoneme sequences suitable for a virtual speech therapy platform. The model has been optimised for practical deployment through methodical training on a high-quality pronunciation dataset, balanced teacher forcing, and a strong evaluation framework that includes both controlled and real-world settings. In particular, the Phoneme Error Rate shows that the model generalises well across clean, curated datasets and user-generated noisy speech inputs. This supports its inclusion in speech therapy programs aimed at improving pronunciation across diverse user groups.

4. Results and Discussions

Here, researchers present the results of audio analysis using different techniques to evaluate the speech of patients undergoing virtual speech therapy. The main objective of this system is to provide Speech-Language Pathologists (SLPs) with a complete set of monitoring tools. By presenting waveforms, spectrograms, and Mel-Frequency Cepstral Coefficients (MFCCs), researchers aim to make these visualizations easier for SLPs to understand, allowing them to evaluate speech qualities such as fluency, clarity, and phoneme accuracy. It is meant to advance SLPs' capacity to deliver focused, data-informed therapy to their patients, especially virtually. This system's outputs are tailored for SLPs to understand and apply in speech therapy sessions. By seeing the waveform, spectrogram, and MFCC graphs, SLPs can obtain insights into various facets of a patient's speech. These representations show speech dynamics, pronunciation accuracy, and fluency, helping the SLP identify specific areas of difficulty. With this information, SLPs can design their therapy interventions to address these difficulties, measure progress, and monitor improvement over time. The system is not just a diagnostic tool but also a way to maximize therapy efficacy

through real-time, actionable feedback. In Figure 3, the waveform graph is a simple yet powerful tool that visually shows the speech signal's amplitude over time. For an SLP, this graph is crucial in assessing the timing and volume of the patient's speech.

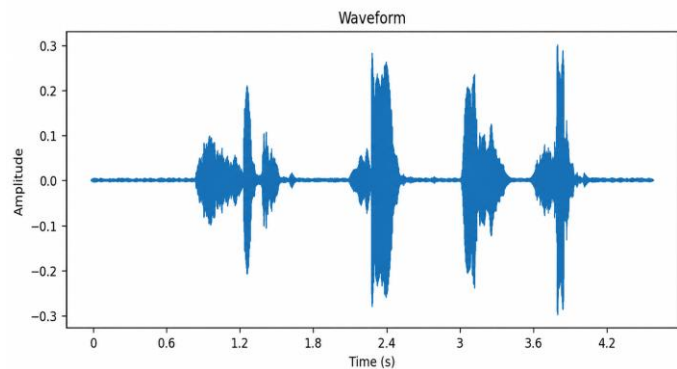


Figure 3: Waveform graph

or erratic pacing—typical traits among patients suffering from fluency disorders such as stuttering. It also enables detection of volume variations or loudness shifts, which can indicate voice disorders. Waveform variations can also identify intonation issues, enabling the SLP to analyze the speech's naturalness and expressiveness. The waveform is therefore a first-stage diagnostic tool, providing rapid visual information to evaluate fluency and volume control.

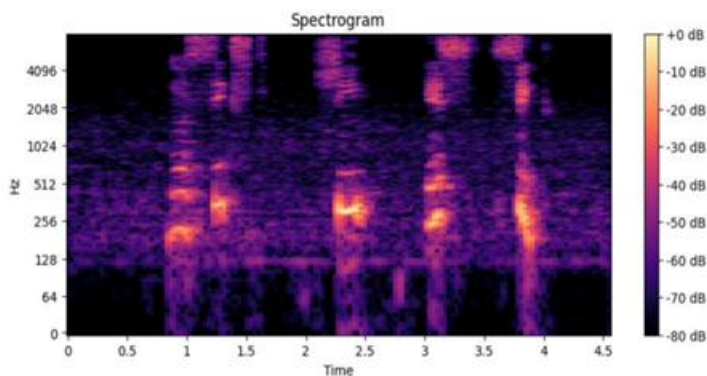


Figure 4: Spectrogram

In Figure 4, the spectrogram provides a more precise display by showing how the speech signal's frequency content changes over time. This instrument is particularly valuable to SLPs when evaluating articulation and phoneme clarity. By examining the spectrogram, an SLP can see temporal aspects of speech sounds, including duration and clarity. This helps detect phoneme misarticulations or timing errors. The spectrogram also highlights changes in pitch and energy distribution, providing additional information about fluency and expressiveness. For example, it can help SLPs determine whether the patient produces clear, distinct sounds or combines or slurs phonemes. It can also detect difficulties with particular phonetic features (e.g., consonants or vowels) and their production, allowing the SLP to modify treatment. In Figure 5, MFCCs (Mel-Frequency Cepstral Coefficients) are a more sophisticated method of analyzing the spectral properties of speech.

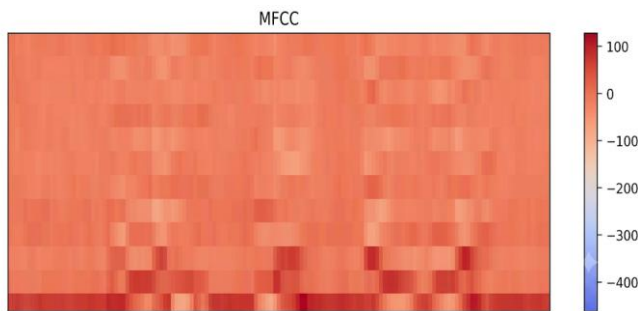


Figure 5: MFCC graph

This graph is particularly useful for phoneme-level analysis. Because MFCCs aim to mimic the human auditory system's sensitivity to sound frequencies, they provide a concise yet in-depth representation of speech characteristics that can identify subtle differences in phoneme production. For an SLP, the MFCC graph is highly valuable for recognizing mispronunciations, particularly when evaluating phoneme accuracy in speech sound disorder patients. By matching the MFCC characteristics of the patient's speech against reference models, the SLP can identify misarticulations, substitutions, or distortions in phoneme production. Furthermore, MFCCs can be employed to determine the Phoneme Error Rate (PER), which offers a quantifiable measure of speech accuracy. This helps SLPs monitor progress and assess the effectiveness of their therapy methods. Integrating the three graphical illustrations—waveform, spectrogram, and MFCCs—gives SLPs a detailed snapshot of a patient's speech dynamics, articulation, and phoneme accuracy. The waveform chart is especially effective for monitoring timing, pacing and volume, while the spectrogram captures detailed insights into phoneme articulation and pitch changes. The MFCC chart, by contrast, provides accurate phoneme-level analysis that helps SLPs catch even subtle pronunciation errors and track improvements in speech accuracy. Together, these tools provide a detailed, multidimensional analysis of speech that enables SLPs to offer personalized, targeted therapy. By regularly monitoring these visualizations, SLPs can adjust interventions as needed. Researchers present the results of the audio analysis performed using various techniques to assess the speech of patients undergoing virtual speech therapy.

The main objective of this system is to equip Speech-Language Pathologists (SLPs) with a complete suite of tools to observe, analyze, and give feedback on their patients' speech. Using waveforms, spectrograms, and Mel-Frequency Cepstral Coefficients (MFCCs), researchers aim to help SLPs interpret these visualizations and analyze speech features such as clarity, fluency, and phoneme accuracy. This system aims to help SLPs provide patients with targeted, data-driven therapy, especially in a virtual environment. The system's output is designed for SLPs to review and apply in speech therapy sessions. With visualizations of the waveform, spectrogram, and MFCC graphs, SLPs can gain insight into various aspects of a patient's speech. Each representation provides information on speech dynamics, pronunciation accuracy, and fluency, enabling the SLP to pinpoint areas of difficulty. With this information, SLPs can adapt their therapy approach to meet these challenges, track progress, and monitor improvement over time. The system is not just a diagnostic tool; it also maximizes therapy effectiveness by providing real-time, actionable feedback. **Waveform Graph:** The waveform graph is a simple yet effective tool that visually displays the speech signal's amplitude over time. For an SLP, this graph is essential to assess the timing and loudness of the patient's speech. From the waveform, SLPs can identify problems like excessive pauses, hesitation, or uneven pacing—typical traits in patients with fluency disorders like stuttering. It also helps identify volume inconsistencies or loudness changes, which may indicate voice disorders. Changes in the waveform can also identify intonation difficulties, enabling the SLP to evaluate the naturalness and expressiveness of speech. The waveform is therefore an initial diagnostic aid, providing rapid visual feedback for evaluating fluency and volume control.

Spectrogram Graph: The spectrogram gives a more detailed representation by showing how the frequency content of the speech signal changes over time. It is especially useful for SLPs assessing articulation and phoneme clarity. From the spectrogram, an SLP can identify temporal features of speech sounds, including duration and intelligibility. This helps identify phoneme misarticulation or timing errors. The spectrogram also shows changes in pitch and energy distribution, providing additional information on speech fluency and expressiveness. For example, it can help SLPs determine whether the patient produces clear, distinct sounds or combines or slurs phonemes. It can also identify difficulties with certain phonetic characteristics (i.e., consonants or vowels) and how they are produced, allowing the SLP to modify treatment accordingly. **MFCC Graph:** MFCCs (Mel-Frequency Cepstral Coefficients) are a more sophisticated method of examining the spectral features of speech. This graph is particularly helpful for phoneme-level analysis. Because MFCCs mimic the human auditory system's frequency sensitivity, they provide a precise, concise representation of speech characteristics that can detect minute differences in phoneme production. The MFCC graph is worth its weight in gold for an SLP when identifying mispronunciations, particularly when evaluating phoneme accuracy in speech-disordered patients. By comparing the MFCC features of a patient's speech to reference models, the SLP can identify misarticulations, substitutions, or distortions in phoneme production. Moreover, MFCCs can be employed to compute the Phoneme Error Rate (PER), yielding a quantifiable measure of speech accuracy. This helps SLPs monitor progress and assess the effectiveness of their therapy methods. **Summary and Significance for SLPs:** Using these three graphical representations together gives SLPs a comprehensive view of a patient's speech, covering speech dynamics, articulation, and phoneme accuracy. The waveform chart helps measure timing, pacing, and loudness, while the spectrogram provides detailed information on phoneme articulation and pitch variations.

Table 1: Reference for sentence indexing used in graphical analysis

No.	Sentence
1	She sells seashells on the seashore
2	The sixth sheep is sick.
3	The thieves thought they stole the throne.
4	A proper cup of coffee from a proper coffee pot

5	Trick trains travelled through tunnels.
6	Valve failures severely damaged their vehicles.
7	Visualize the zenith through thick fog.
8	Zebras have very thin stripes.
9	They gave me a vase of azaleas today.

The MFCC chart provides high-precision, phoneme-level analysis, enabling SLPs to identify subtle pronunciation errors and monitor improvements in speech accuracy. Together, these tools offer an in-depth, multidimensional view of speech that allows SLPs to deliver tailored, targeted therapy. With frequent review of these visualizations, SLPs can revise therapy protocols based on the patient's development, pinpoint specific areas of difficulty, and ensure the patient is visibly improving their speech. These graphs serve not only as diagnostic tools but also as a means to enhance the overall effectiveness of virtual speech therapists and their therapy strategies based on the patient's progress, and to identify specific problems. To keep the visualizations compact and unobtrusive, numeric labels on the x-axis of different graphs denote individual sentences, as shown in Table 1. It provides a separate reference that translates these numeric labels back to their corresponding sentence text. This keeps the graphs compact and readable without compromising readers' ability to cross-reference each data point to its corresponding sentence. Table 1 is essential for detailed analysis because it lets researchers interpret error patterns, model predictions, or performance metrics with direct reference to a sentence's linguistic content.

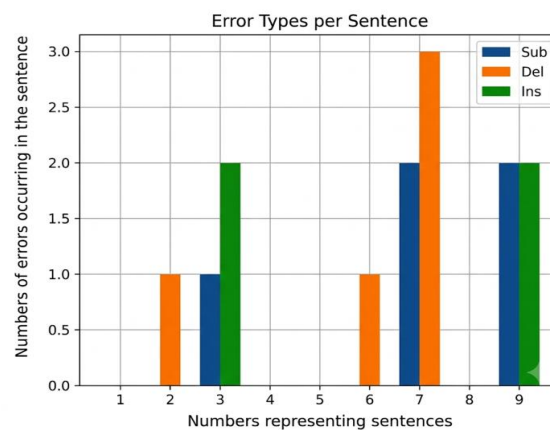


Figure 6: Types of error in sentences

The error analysis offered within the system is informative for speech-language pathologists (SLPs) treating patients remotely. In Figure 6, the information shows specific phoneme-level errors in practice sentences, such as substitutions (e.g., "ick fog" for "thick fog"), deletions, and insertions, with severity ratings from 0.0 (ideal production) to 3.0 (severe difficulty). This detailed error analysis allows SLPs to pinpoint specific articulation difficulties, such as trouble with /θ/ in "throne" or /s/ clusters in "ashore." The system's automatic longitudinal tracking of error patterns makes it easier for clinicians to quantify progress, even in distance settings. By identifying the most difficult phoneme combinations (such as /tr/ in "throne" or vowel sequences in "atrides"), the technology helps SLPs create focused intervention plans.

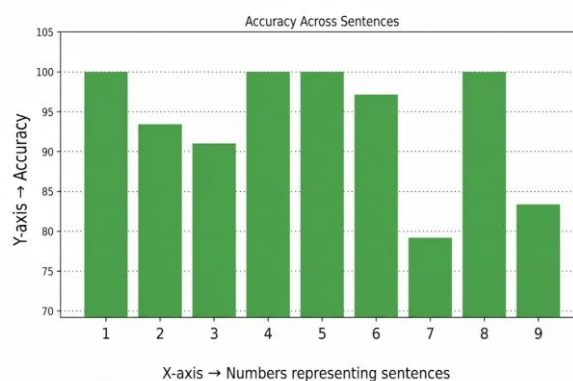


Figure 7: Sentence-wise accuracy

Quantitative error scoring (1.0-3.0) provides a standardized way to prioritize therapy. The results show certain phoneme-level errors in practice sentences, such as substitutions (e.g., "ick fog" instead of "thick fog"), deletions, and insertions, with severity ratings from 0.0 (excellent production) to 3.0 (severe difficulty). This detailed error profiling allows SLPs to pinpoint specific articulation difficulties, like /θ/ trouble in "throne" or /s/ clusters in "ashore." The system's automated monitoring of error patterns over time lets clinicians objectively quantify improvement, even during remote therapy. By identifying the most difficult phonemic combinations (such as the /tr/ sound in "throne" or vowel shifts in "atripes"), the technology helps SLPs craft focused intervention strategies. The quantifiable error rating (1.0-3.0) provides a standardized way to rank therapy focus areas. At the same time, sentence-level analysis captures context-dependent speech difficulty that individual word practice may miss. This data-based methodology augments standard auditory-perceptual assessment procedures, giving SLPs an instrument to generate evidence-based clinical judgments, tailor home practice materials, and document therapeutic gains for clients with articulation disorders or phonological processing impairments.

Sentence-level analysis identifies contextual speech difficulties that isolated word practice may overlook. This data-driven methodology improves on traditional auditory-perceptual evaluation methods by giving SLPs a way to make evidence-based clinical decisions, tailor home practice materials, and document therapeutic results for clients with articulation disorders or phonological processing deficits. The system's sentence-level accuracy ranges from 70% to 95%, showing its ability to detect and quantify speech production difficulties across diverse phonetic environments. As shown in Figure 7, mispronunciations such as "The seabine" (presumably "submarine") and "used the trione" (perhaps "throne") indicate ongoing trouble with /s/-blends and /θ/ (voiced "th") sounds. At the same time, greater accuracy in easier phrases ("per coffee pot") attests to anticipated mastery with frequent phonemes. The gradient decrease in accuracy for compound targets ("ough thick fog" for "through thick fog") helps SLPs set therapy focus areas, such as fricative clusters or rhotic vowels. By mapping these patterns systematically, the tool gives SLPs objective benchmarks to monitor patient improvement, adjust intervention intensity, and adapt home practice (e.g., highlighting /θ/ in words like "their vehicles"). areas, and ensure that the patient is making measurable improvements in their speech (Table 2).

Table 2: Sample of generated phoneme analysis report

Sentence	Phoneme Error Rate (%)	Accuracy (%)
She sells seashells on the seashore	0.00	100.00
The sixth sheep is sick	6.67	93.33
The thieves thought they thrilled the throne	9.09	90.91
A proper cup of coffee from a proper copper coffee pot	0.00	100.00
Tricky trains travel through tunnels	0.00	100.00
Valve failures severely damaged their vehicles	2.94	97.06
Visualize the zenith through thick fog	20.83	79.17
Zebras have very thin stripes	0.00	100.00
They gave me a vase of azaleas today	16.67	83.33

The suggested voice therapy system is based on a Sequence-to-Sequence (Seq2Seq) model implemented with LSTM networks to transform English words into corresponding phoneme sequences. The system is designed to help users recognize and exercise phonemes that are challenging for them by utilizing a character-to-phoneme mapping pipeline. The model shows a strong ability to learn complex correlations between character input sequences and corresponding phoneme outputs. During training, the model regularly decreases its loss, showing that it reduces prediction error efficiently and learns patterns in phoneme representation. This learning behavior is particularly noteworthy given the data source. The model was trained on the CMU Pronouncing Dictionary, a standardized database that includes mappings for more than 134,000 English words and their phonetic transcriptions. By pre-processing the data to include only words with a single pronunciation, the system could focus on learning deterministic mappings without ambiguity and converge more robustly. The embedding and LSTM layers enabled the model to learn an internal representation of how certain character patterns map to phoneme structures, which is important for correct phoneme prediction.

Table 3: Performance metrics of the model

Metric	Value
Accuracy	0.7018
Precision (Macro)	0.5447
Recall (Macro)	0.5075
F1 Score (Macro)	0.5163
Precision (Weighted)	0.6999

F1 Score (Weighted)	0.7018
Recall (Weighted)	0.7018

Table 3 shows the proposed model's performance metrics. A key contributor to the overall system's performance is its reliance on the Google Speech Recognition API, which transcribes user-spoken audio into text. This text is used as input to the phoneme prediction model. The quality of the transcribed text, therefore, affects the accuracy of the final phoneme output. Misrecognitions during this phase—due to ambiguous pronunciation, ambient noise, or variation in accent—have the effect that wrong sequences of characters get supplied to the model, regardless of the model having internalized correct patterns for phonemes. The Phoneme Error Rate (PER) was employed to measure how well the predicted phonemes aligned with the ground truth, in terms of edit distance. The model performed reasonably well on training words and generalized well to unseen but phonetically regular words. However, errors appeared in predictions for user-supplied custom inputs, where phoneme substitutions or deletions sometimes occurred. These errors underscore the need for improved generalization and additional fine-tuning, especially for novel word patterns. These graphs serve not only as diagnostic tools but also as a means to enhance the overall effectiveness of virtual speech therapy.

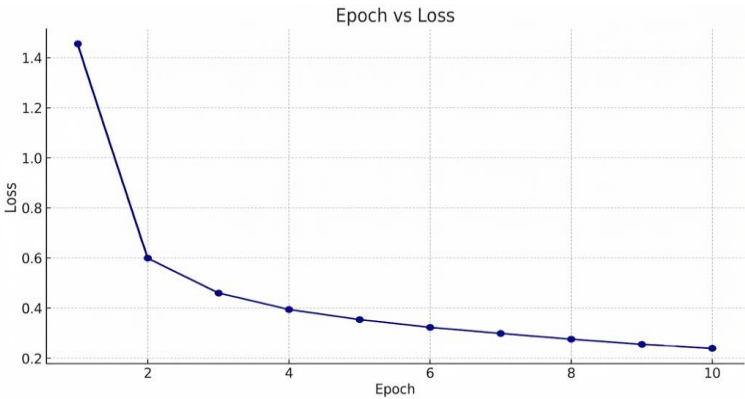


Figure 8: Epoch VS Loss graph during training of seq-to-seq encoder-decoder model

Figure 8 shows the loss vs. epoch graph, which illustrates the training dynamics and convergence behavior of the sequence-to-sequence phoneme prediction model across 10 epochs. The loss curve consistently decreases from 1.4 in the first epoch to 0.2 by the last epoch, reflecting efficient learning and parameter optimization. This steady decrease in loss indicates that the model learned to map input text sequences to their corresponding phoneme outputs, with the most dramatic improvements in the initial training phase (epochs 2-4). The smooth flattening of the curve in subsequent epochs (6-10) suggests the model was reaching convergence, with decreasing returns in loss reduction as it improved its predictions. This training course has significant implications for clinical use in speech-language pathology. The initial steep decline in loss suggests the model rapidly acquired simple grapheme-to-phoneme mappings, consistent with how patients tend to show rapid gains when first exposed to new phonemes during therapy. The more gradual, stable gains in subsequent epochs suggest the model's (and by extension, a patient's) move toward learning more difficult phonological patterns. For SLPs, this implies that early therapy sessions may be most effective at generating obvious gains. Still, for more subtle articulation skills, practice must be continuous—consistent with motor-learning principles in speech rehabilitation. The final loss value of 0.2 indicates strong model performance, with the clinical implication that the system can offer precise, real-time feedback in speech therapy drills. Technically, the lack of spikes or plateaus in the loss curve indicates stable training with minimal overfitting or optimization issues. Such reliability is paramount for clinical use, as it guarantees consistent results when processing patient speech. Selecting 10 epochs balances computational cost with model performance, making the system feasible for practical use without a long training period.

For subsequent versions, the curve suggests that additional epochs may yield little further loss reduction. Still, methods such as learning rate scheduling or data augmentation may further improve fine-grained phoneme discrimination—especially for difficult sounds such as /θ/ or consonant clusters. Overall, the epoch vs. loss graph confirms the model's learning effectiveness and parallels therapeutic improvement in speech rehabilitation. The evident downward trend confirms the system's clinical readiness. At the same time, the training dynamics provide insight into machine learning optimization and the anticipated direction of patient progress in phoneme acquisition. Both views strengthen the case for incorporating such models into SLP practice, since they work reliably and reflect principles consonant with therapeutic realities. In Figure 9, the time per epoch versus epoch graph reveals interesting information about the computational cost and stability of the model's training procedure. Over 10 epochs, the time per epoch is highly uniform, ranging from 44.2 to 45.8 seconds. This small variation indicates a stable, well-optimized training process in which each epoch has a comparable computational cost. The mild downward slope in time per epoch (from about 45.5 seconds in early epochs to around 44.5 seconds in later epochs) suggests modest gains in processing

efficiency over training, perhaps due to more efficient memory use or hardware warm-up effects. This stability in training time has significant practical consequences for clinical deployment. The consistent length of each training epoch allows SLPs and researchers to accurately predict the time needed for model updates or retraining on new patient data. The effective timing (less than 46 seconds per epoch) makes the system feasible for routine updates and adjustments to individual patient requirements without requiring excessive computational resources or time.

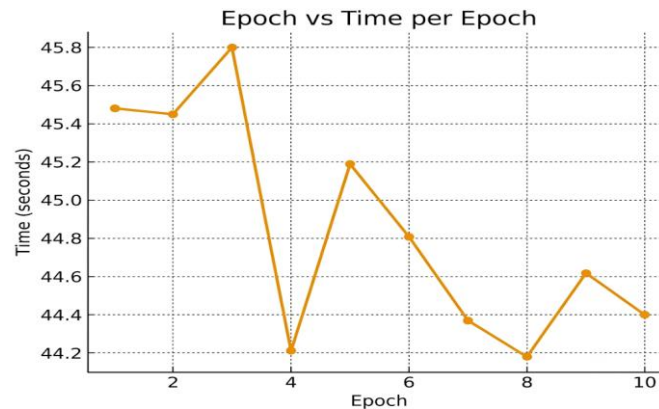


Figure 9: Epoch VS Time per epoch during training of seq-to-seq encoder-decoder model

This is especially useful in clinical environments where therapy methods may need to change quickly as patients improve. The stability exhibited in this plot also speaks positively about the overall system design. The lack of large time variances shows that the data pipeline is well-balanced during training, with no bottlenecks forming in data loading or processing. For SLPs using this technology, this means a consistent system they can reuse across various clinical situations. The timely efficiency also means the model can be retrained or fine-tuned within an available therapy session if needed, enabling real-time adjustments to meet specific patient needs. From a research perspective, the timing results confirm that the model architecture is computationally efficient while retaining training performance, as reflected in the loss curve. The intersection of stable timing and diminishing loss (as shown in the graph above) suggests a balance between model complexity and practical usefulness. This makes the system especially appropriate for clinical settings, where accuracy and responsiveness are both crucial in therapy applications. The consistent timing pattern between epochs also supports replicable training, a key requirement for evidence-based clinical devices. Although the model has demonstrated good learning ability, improvement is still possible. Additional fine-tuning of the network, testing alternative attention mechanisms, and covering more vocabulary to support multiple-pronunciation words would go a long way toward making the model more robust. Additionally, replacing or complementing the speech recognition module with a more configurable alternative may reduce transcription errors and improve end-to-end performance. Incorporating user feedback loops to confirm or correct phoneme outputs may also help the system adapt to individual learning needs, making it more effective in therapy and language-learning applications. Overall, the model successfully translates English text into phonemes in an organized and sensible manner. While its isolated-condition performance is promising, the entire pipeline's accuracy depends on several components—most notably the speech-to-text component.

5. Conclusion

In conclusion, the AI-powered voice therapy system developed in this paper represents a significant advancement in speech-language therapy. By leveraging cutting-edge phoneme recognition, attention-based models, and real-time pronunciation scoring, the system provides users with consistent, high-quality pronunciation feedback that mimics the in-clinic experience. Its ability to offer real-time, adaptive exercises tailored to the individual's skill level, coupled with telepractice integration, enhances therapy flexibility, making it more accessible and convenient for patients while reducing logistical challenges. This model not only supports independent practice but also equips clinicians with advanced tracking and analytics, enabling remote progress monitoring and timely interventions. Furthermore, motivational features such as gamification, milestone tracking, and offline support increase patient compliance, making therapy more engaging and rewarding. The system's integration of AI-driven tools and therapist-guided supervision creates a unique hybrid model that supports continuous improvement without overwhelming clinicians. This paper has shown that combining AI's precision with human expertise creates an effective, scalable solution that transforms how voice therapy is delivered. Looking forward, the system is poised for further refinement, with potential for greater personalization, expanded features, and integration with additional therapeutic modalities. As technology and AI continue to evolve, this platform will play an increasingly critical role in making speech therapy more personalized, accessible, and efficient for diverse patient populations. Ultimately, the AI-driven voice therapy system

exemplifies a future in which patients and clinicians alike can harness technology to achieve better therapeutic outcomes with greater convenience and effectiveness.

Acknowledgment: N/A

Data Availability Statement: The data for this study are available upon request from the corresponding author.

Funding Statement: This manuscript and research paper were prepared without any financial support or funding.

Conflicts of Interest Statement: The authors have no conflicts of interest to declare. This work represents a new contribution by the authors, and all citations and references are appropriately included based on the information utilized.

Ethics and Consent Statement: This research adheres to ethical guidelines and obtains informed consent from all participants. Confidentiality measures were implemented to safeguard participant privacy.

References

1. M. Orosoo, N. Raash, M. Treve, H. F. M. Lahza, N. Alshammry, J. V. N. Ramesh, and M. Rengarajan, "Transforming English language learning: Advanced speech recognition with MLP-LSTM for personalized education," *Alexandria Engineering Journal*, vol. 111, no. 1, pp. 21–32, 2025.
2. M. A. Sadik, A. S. ElSayed, S. M. Moussa, and Z. T. Fayed, "2D spectrogram analysis using vision transformer to detect mispronounced Arabic utterances for children," *Applied Soft Computing*, vol. 165, no. 11, p. 112052, 2024.
3. M. Shahin, J. Epps, and B. Ahmed, "Phonological level wav2vec2-based mispronunciation detection and diagnosis method," *Speech Communication*, vol. 173, no. 9, pp. 1–17, 2025.
4. C. Zhu, A. Wumaier, D. Wei, Z. Fan, J. Yang, H. Yu, Z. Kadeer, and L. Wang, "Pronunciation error detection model based on feature fusion," *Speech Communication*, vol. 156, no. 1, pp. 1–12, 2024.
5. T. Khodaveisi, H. Bouraghi, S. Saeedi, M. Ghazisaeedi, M. S. Seifpanahi, S. Ahsanifar, and S. Vafaeeeyan, "Mobile health in communication disorders: Willingness to use, attitude, advantages, and challenges from the perspective of patients," *BMC Health Services Research*, vol. 25, no. 1, pp. 1–15, 2025.
6. I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *NIPS'14: Proceedings of the 28th International Conference on Neural Information Processing Systems*, Montreal, Canada, 2014.
7. R. Lakshminarayanan, A. Shaik, and A. Balasundaram, "Automated speech therapy through personalized pronunciation correction using reinforcement learning and large language models," *Results in Engineering*, vol. 25, no. 3, pp. 1–15, 2025.
8. H. Nam and Y. H. Park, "Coherence-based phonemic analysis on the effect of reverberation to practical automatic speech recognition," *Applied Acoustics*, vol. 227, no. 1, pp. 1–9, 2025.
9. H. Garg, S. Jhunthra, M. Kindra, V. Dixit, and V. Gupta, "A deep learning-based integrated voice assistance system for partially disabled people," in *Uncertainty in Computational Intelligence-Based Decision Making: Advanced Studies in Complex Systems*, Academic Press, Massachusetts, United States of America, 2025.
10. Y. Geng, "Design of English teaching speech recognition system based on LSTM network and feature extraction," *Soft Computing*, vol. 28, no. 6, pp. 13873–13883, 2024.
11. D. Ji, "The impact of transformer-based chatbots on second language speaking competence in blended learning," *Discover Computing*, vol. 29, no. 3, pp. 1–31, 2026.
12. F. Esmailzadeh, S. Ghahari, and G. Rohani, "Artificial intelligence in language teaching and assessment: A systematic review of research (2020–2024)," *The Asian Journal of Applied Linguistics*, vol. 9, no. 1, pp. 1–23, 2025.
13. E. I. Toki, G. Tatsis, V. A. Tatsis, K. Plachouras, J. Pange, and I. G. Tsoulos, "Employing Classification Techniques on Smartspeech Biometric Data towards Identification of Neurodevelopmental Disorders," *Signals*, vol. 4, no. 2, pp. 401–420, 2023.
14. Ş. S. Çalık, A. Küçükmanisa, and Z. H. Kilimci, "A novel framework for mispronunciation detection of Arabic phonemes using audio-oriented transformer models," *Applied Acoustics*, vol. 215, no. 1, p. 109711, 2024.
15. A. Bhardwaj, M. Sharma, S. Kumar, S. Sharma, and P. C. Sharma, "Transforming pediatric speech and language disorder diagnosis and therapy: The evolving role of artificial intelligence," *Health Sciences Review*, vol. 12, no. 9, pp. 1–7, 2024.

Publisher's Note: The publisher remains impartial concerning jurisdictional claims in published maps and institutional affiliations. Responsibility for the content rests entirely with the authors and does not necessarily reflect the publisher's perspectives.