Review

AI-Powered Speech Therapy Assistant

Abhinav M Biju *, Joyal James *, Nahil Zubair Ridwan *, T S Vishal Menon *, Surya S, Sulaja Sanal, Sabeena K

Posted Date: 24 November 2025

doi: 10.20944/preprints202511.1793.v1

Keywords: AI-powered therapy; speech sound disorders; automatic speech recognition; natural language processing; conversational ai; gamified therapy; assistive technology

 doi:10.20944/preprints202511.1793.v1

Article

AI-Powered Speech Therapy Assistant

Abhinav M Biju *, Joyal James *, Nahil Zubair Ridwan *, T.S Vishal Menon *, Surya S, Sulaja Sanal and Sabeena K

Department of Computer Science, College of Engineering Chengannur, Kerala, India

* Correspondence: abhinavmbiju@gmail.com (A.M.B.); joyalchandrankunnel@gmail.com (J.J.); nahilpp@gmail.com (N.Z.R.); tsvishalmenon2003@gmail.com (T.S.V.M.)

Abstract

AI-Powered Speech Therapist Assistant is an intelligent software solution designed to support individuals with Speech Sound Disorders (SSD) and related communication challenges by automating key aspects of speech-language therapy. Built using advanced Artificial Intelligence (AI) techniques such as Automatic Speech Recognition (ASR), Natural Language Processing (NLP), and generative conversational models, the system provides personalized therapy sessions, pronunciation feedback, and interactive learning exercises in real time. Different functional modules—such as speech recognition, pronunciation analysis, gamified therapy tasks, therapist dashboards, and conversational practice—work together to deliver a comprehensive therapeutic experience. For instance, the pronunciation analysis module evaluates articulation, stress, and rhythm, while the conversational agent simulates real-life dialogue to improve expressive and receptive language skills. Progress is continuously tracked, and updates are reflected across modules to ensure adaptive feedback and personalized learning pathways. Key features of the system include impaired-speech dataset integration, real-time feedback, gamified exercises, role-based access for therapists and patients, and a secure data management layer. The software ensures accessibility, engagement, and consistency, while reducing therapist workload through automation. By leveraging AI technologies, the system offers a cost-effective and scalable solution for schools, rehabilitation centers, and home-based therapy, making professional speech support more widely accessible. Future enhancements include multilingual support, cloud-based telepractice, advanced analytics dashboards, emotion-aware conversational agents, and clinical integration with healthcare systems.

Keywords: AI-powered therapy; speech sound disorders; automatic speech recognition; natural language processing; conversational ai; gamified therapy; assistive technology

1. Introduction

Speech therapy is a critical domain in healthcare and education, supporting individuals with Speech Sound Disorders (SSD), articulation difficulties, and other communication challenges. The primary goal of speech therapy is to improve clarity, comprehension, and expressive ability, thereby enhancing overall communication skills and quality of life. Effective therapy ensures that patients develop the ability to produce correct sounds, construct sentences, and engage confidently in social, academic, and professional interactions. However, accessibility, affordability, and limited availability of Speech-Language Pathologists (SLPs) remain major barriers, especially in rural and underserved communities [5].

Traditional approaches to speech therapy often involve face-to-face sessions with trained professionals, utilizing manual techniques such as flashcards, repetition drills, and auditory exercises. While these methods are effective, they are time-consuming, resource-intensive, and difficult to scale as the demand for speech therapy grows. In many regions, waiting lists for SLP services extend for months, leaving patients without timely support [6]. Recent advancements in artificial intelligence (AI) have

 doi:10.20944/preprints202511.1793.v1

created opportunities to address these challenges by offering scalable, adaptive, and cost-effective therapy solutions.

The emergence of AI-driven technologies has transformed speech therapy into a domain where automation and personalization coexist. Key innovations include Automatic Speech Recognition (ASR), Natural Language Processing (NLP), and conversational AI. These technologies enable real-time pronunciation assessment, personalized feedback, and interactive therapy tasks that adapt to the progress of the individual [7]. Similar to how Just-in-Time (JIT) optimized efficiency in manufacturing, AI-powered therapy introduces precision and adaptability into healthcare by closely aligning therapy activities with individual needs while minimizing resource waste.

In recent years, digital health environments have fostered the evolution of automated speech therapy systems, incorporating gamification, mobile applications, and online platforms. These technologies provide engaging therapy experiences, reduce human error, and offer continuous monitoring of progress. AI tools not only track articulation and pronunciation accuracy but also analyze suprasegmental features such as rhythm, intonation, and stress, thereby offering a holistic view of the patient's communication abilities [4].

Beyond direct patient support, AI systems can also enhance the role of therapists by acting as intelligent assistants. Features such as patient monitoring dashboards, task assignment modules, and performance tracking tools allow therapists to oversee multiple patients simultaneously, customize therapy plans, and intervene when needed. This dual focus—empowering patients with self-practice tools while equipping therapists with digital support systems—ensures a collaborative and more effective therapeutic process. By combining patient-facing and therapist-assisting features, AI platforms bridge the gap between accessibility and professional oversight, creating a balanced ecosystem for communication rehabilitation.

Despite significant progress, challenges remain. Integrating AI-powered speech therapy systems with existing clinical practices raises issues of accuracy, ethical data handling, and inclusivity across diverse languages and dialects. Although digital therapy platforms aim to provide holistic solutions, ensuring patient trust and compliance requires careful collaboration between technologists and healthcare professionals [2].

Examining current AI-driven speech therapy tools highlights the unique needs of different populations, ranging from children with articulation disorders to adults recovering from stroke or traumatic brain injury. Small-scale mobile applications may suffice for home practice, while institutional deployments demand robust, clinically validated systems. Understanding these differences is essential for guiding the future of AI-powered therapy. As society continues to adapt to evolving healthcare demands, the development of intelligent, accessible, and adaptable therapy tools becomes increasingly critical.

1.1. Technological Advances in Speech Therapy

Over the past decade, technological innovations have expanded the scope of speech therapy from conventional in-person sessions to AI-assisted, interactive platforms. ASR and NLP technologies now allow therapy systems to not only detect mispronunciations but also provide real-time corrective feedback and performance tracking. These advancements represent a paradigm shift, enabling systems that can analyze, adapt, and support therapy beyond human limitations, thereby enhancing accessibility and efficiency [5].

1.2. Incorporating Real-Time Feedback

The proposed AI-Powered Speech Therapist Assistant emphasizes real-time analysis and updates to enhance therapeutic outcomes. By instantly detecting errors in articulation, stress, and rhythm, the system empowers patients to self-correct during practice. This immediate feedback mechanism ensures faster learning and greater confidence in communication. Moreover, data is synchronized across modules, allowing therapists to monitor patient progress and tailor sessions accordingly, thus fostering a collaborative and adaptive therapy environment [6].

 doi:10.20944/preprints202511.1793.v1

1.3. Portable and Cost-Effective Implementation

Implemented as a software platform accessible via desktop and mobile devices, the system is designed to be affordable, lightweight, and widely deployable. By leveraging AI models trained on impaired-speech datasets, it supports diverse patient needs with high accuracy. The system integrates modules for pronunciation assessment, gamified exercises, conversational practice, therapist dashboards for patient monitoring, and task assignment features, ensuring comprehensive support for both patients and clinicians. Cost-effectiveness, coupled with its scalability, makes the solution suitable for clinics, schools, and home-based therapy.

This paper discusses the design, implementation, and evaluation of the proposed AI-Powered Speech Therapist Assistant while reviewing current literature that underscores the importance of AI technologies in transforming therapeutic practices. By combining real-time speech analysis, adaptive exercises, conversational AI, and therapist-assisting tools, the system marks a significant step forward in digital healthcare solutions. Furthermore, it highlights the potential for AI-driven systems to improve accessibility, efficiency, and patient engagement in speech therapy. Moving forward, enhancements such as multilingual support, cloud synchronization, and integration with healthcare platforms can further elevate the system's impact on communication rehabilitation.

2. Literature Review

The literature on AI in speech-language pathology (SLP) highlights both opportunities and challenges in integrating emerging technologies into clinical practice. Prior studies demonstrate that AI and machine learning can support assessment and intervention through tools like computer vision for articulatory analysis, robotic assistants, serious games, and automated feedback systems. Recent work emphasizes the potential of generative AI, such as ChatGPT and DALL·E 2, to assist SLPs in creating therapy materials, translating resources, and simulating human-like interactions, thereby saving time and reducing administrative burdens. However, researchers caution against overreliance on AI in client-facing care, citing ethical concerns around privacy, algorithmic bias, and the irreplaceable value of human empathy in therapy. Overall, existing research suggests that while AI can streamline documentation, scheduling, and resource preparation, its responsible adoption requires co-design with SLPs to ensure tools enhance capacity and job satisfaction without compromising quality of care [5].

Recent studies on Arabic Automatic Speech Recognition (ASR) have analyzed various architectures, including HMM, DNN, RNN, sequence-to-sequence, and hybrid models, alongside feature extraction, acoustic, and language modeling techniques. Challenges were identified as language-dependent, such as Arabic grammar, diacritics, and dialects, and language-independent, including noise, speaker variability, and devices. Only nine publicly available Arabic ASR corpora were identified, including MGB2–MGB5, the Quranic corpus, Arabic Digits, and Tunisian MSA corpus, with dialectal corpora emerging but limited. Deep learning models such as CNN-LSTM, Bi-LSTM with attention, and DeepSpeech2 outperformed traditional HMMs, achieving WERs of 12.5% (MGB2), 27.5% (MGB3), and 33.8% (MGB5), though still below human-level accuracy. Hybrid word-character recognition improved handling of unknown words. Applications include speech-to-text, speaker recognition, education, and healthcare, but limitations remain, including dataset scarcity, heavy dialectal variation, small corpora, and WERs above human performance. These findings highlight both the progress and challenges of Arabic ASR for practical applications [6].

Based on recent studies utilizing the AphasiaBank dataset, researchers have explored the fine-tuning of transformer-based language models, particularly BERT, for supporting speech and language rehabilitation in individuals with aphasia. The methodology involved preprocessing transcripts by removing fillers, gestures, and redundancies, followed by fine-tuning BERT with masked language modeling at both sentence and paragraph levels. In addition to contextual prediction tasks, BERT was further adapted for question-answering, enabling deeper semantic understanding of impaired speech. The AphasiaBank corpus, which spans 12 languages and includes transcripts, audio, and video data

 doi:10.20944/preprints202511.1793.v1

with detailed morphological-syntactic annotations, provided a diverse foundation covering multiple aphasia types and speech impairments. Results demonstrated that BERT generated contextually accurate and fluent predictions, with moderate inter-rater agreement (Kappa = 0.32, correlation = 0.61–0.74), and the integration of question-answering tasks improved contextual interpretation. However, limitations were noted, including reliance on a single dataset, subjectivity in manual evaluations, limited objective performance metrics, and the high computational demands of fine-tuning. These findings highlight both the promise and the challenges of applying large language models in automated speech and language therapy research [7].

Research on pronunciation assessment has compared traditional human-rated methods, which often suffer from bias, inconsistency, and subjectivity, with AI-powered platforms that employ automatic speech recognition (ASR) and speech analysis tools. Modern systems, such as SpeechRater in TOEFL and automated scoring in PTE, leverage large speech corpora, Hidden Markov Models, acoustic models, and n-gram language models to provide scalable and consistent evaluations. These AI-driven tools have proven particularly valuable in large-scale testing environments like IELTS, TOEFL, PTE, and TOEIC, offering rapid, objective, and cost-effective scoring while also delivering detailed, adaptive feedback at the phoneme and utterance level. By prioritizing suprasegmental features such as stress, rhythm, and intonation, these systems enhance speech intelligibility and learning outcomes. Nonetheless, challenges persist, including bias and validity concerns, limited representation of diverse accents, privacy and data ownership risks, and ethical debates around fairness and accountability in high-stakes testing. Over-reliance on automated assessments also risks overlooking the communicative context and nuanced judgment provided by human evaluators [4].

Recent work on child speech synthesis has applied transfer learning to adapt pretrained adult text-to-speech (TTS) models for children's voices. By fine-tuning adult-trained models such as Tacotron within a multi-speaker TTS pipeline—including a speaker encoder, acoustic model, and vocoder—researchers created child-like synthetic speech using both adult and child datasets. The MyST corpus of 393 hours of child speech was partially cleaned and trimmed to produce the 19.2-hour TinyMyST dataset, supplemented with adult corpora like LibriSpeech, VoxCeleb, and VCTK to improve generalization. Evaluation using subjective Mean Opinion Scores (MOS), objective MOSNet, ASR word error rates, and speaker similarity testing demonstrated promising results, with synthetic voices closely matching real child speech in intelligibility (MOS 4.1) and naturalness (MOS 3.9). However, challenges remain, including occasional word distortions, mispronunciation of certain phonemes, limited availability of high-quality child speech data, and forced alignment issues due to adult-trained models. These findings highlight the potential of transfer learning for child-specific TTS while underscoring the need for cleaner, more representative datasets [2].

Recent work has investigated the effectiveness of personalized Automatic Speech Recognition (ASR) models for individuals with speech disorders using data from Project Euphonia. The study compared unadapted, speaker-independent RNN-T models with personalized, speaker-dependent models trained on approximately 10 hours of speech per participant, incorporating SpecAugment for robustness. Speech data included both read and conversational utterances collected from 27 US and UK participants across 13 etiologies, ranging from mild to severe impairment, with recordings captured in real-world environments on personal devices. Results showed that personalized models achieved substantial improvements, reducing median word error rate (WER) to 10.3 compared to 60.8 for unadapted models. Recognition accuracy varied by speech type, with higher WER for conversational speech (36.1) than for read speech (14.6), though training with conversational or mixed read+conversational data enhanced performance. Limitations included a relatively small and etiologically imbalanced participant pool, restrictions on data availability due to privacy concerns, and a focus on English-only datasets. Additionally, the evaluation centered primarily on WER, leaving semantic accuracy and user experience underexplored. Overall, the study underscores the promise of personalized ASR for disordered speech while highlighting the need for more diverse datasets, cross-linguistic validation, and broader evaluation metrics [8].

 doi:10.20944/preprints202511.1793.v1

Jelassi et al. (2024) developed an AI-driven online therapy platform by adapting the QuartzNet 15x5 automatic speech recognition (ASR) model for French and integrating it with the Rasa NLP framework to enable conversational interactions. The system was trained on the Mozilla Common Voice French dataset, enriched with diverse accents, age groups, dialects, and background noise to simulate clinical and home environments, while metadata provided contextual robustness. Evaluation showed that the ASR achieved a 14% word error rate, and the NLP system reached an F1-score of 62.7%, with key tasks such as scheduling appointments exceeding 90% accuracy.

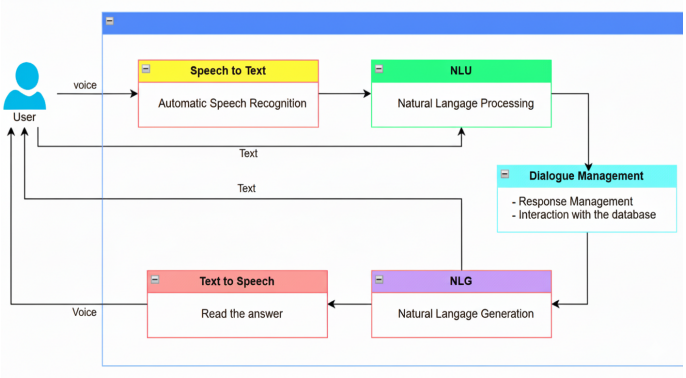

Figure 1. Voice Assistant Flowchart.

The study demonstrated that AI systems can effectively support therapy-related interactions, but challenges remain in handling casual speech, humor, sarcasm, and underrepresented dialects, as well as ethical concerns related to bias and sensitive data privacy [9].

Zeeshan, Bogue, and Asghar (2025) explored the applicability of diverse automatic speech recognition (ASR) platforms for psychiatric treatment session transcription through a two-stage process. In the first stage, they reviewed 32 available ASR and speech-to-text tools for suitability in psychiatric contexts, and in the second stage, they tested 9 shortlisted tools on publicly available psychiatric audio clips. The evaluation considered word error rate (WER), diarization error rate (DER), and inference time, with results showing generally strong transcription accuracy (WER 0–7%) but significant variability in diarization performance (DER 2–32%). Errors were especially pronounced in same-gender speaker interactions, highlighting challenges in distinguishing overlapping voices. The study further compared synchronous and asynchronous diarization approaches, noting that while synchronous diarization better safeguarded privacy, it lacked stability compared to asynchronous methods. Despite these promising insights, the work faced notable limitations, including the absence of large-scale psychiatric datasets due to ethical and privacy concerns, risks of privacy breaches in asynchronous methods, and reduced accuracy in synchronous approaches. Moreover, because testing was limited to controlled English recordings with minimal background noise, the generalizability of the findings to real-world psychiatric settings remains constrained [11].

A recent systematic review examined approaches for detecting dysarthria severity levels using AI and machine learning models, screening 978 studies and identifying 44 as relevant. The review compared traditional clinical assessments by speech-language pathologists with automated methods leveraging audio, image, video, and text features. Among the datasets, TORGO, UA Speech, Qolt, and Nemours were frequently used, though data scarcity, imbalance, and limited language diversity were common challenges. Results indicated that audio-based features remain the dominant input, achieving accuracies up to 96%, while deep learning models such as CNNs and LSTMs performed strongly with accuracies between 95–99%. Moreover, multimodal approaches that combined audio and image features reached up to 99.5% accuracy, highlighting their potential for comprehensive assessment. However, the review emphasized limitations, including small and imbalanced datasets, difficulty in classifying mild dysarthria, gender and language biases, and the high computational cost of deep learning models, which restrict scalability and generalizability [10].

 doi:10.20944/preprints202511.1793.v1

This study presents a speech-to-text and text summarization system that integrates Google's API with custom preprocessing and NLTK-based summarization techniques. Speech was recorded via microphone and initially transcribed by the API, after which punctuation markers were inserted to improve clarity and processing. The system handled various input sources, including recorded speech, long documents, and website text, using sentence and word tokenization along with word frequency–based ranking to generate summaries. Compared to the Gensim summarization library, the proposed model produced faster, more flexible summaries that preserved meaning rather than reducing outputs to single lines. Despite these strengths, the approach remains limited by its support for only basic punctuation, reliance on pause detection for quality, lack of multilingual evaluation, and minimal benchmarking beyond Gensim [1].

The AUTO-AVSR framework was introduced for large-scale audio-visual speech recognition (AV-ASR) to improve robustness against acoustic noise while minimizing manual labeling requirements. It leverages automatically generated transcriptions from large unlabelled datasets using pre-trained ASR models such as Whisper, wav2vec 2.0, HuBERT, and Conformer-Transducer. By integrating labelled datasets (LRS2 and LRS3) with unlabelled sources (AVSpeech and VoxCeleb2), the study achieved over 3,448 hours of combined training data. The model employs a ResNet-Conformer-Transformer architecture, where 3D/2D ResNet-18 encoders process the visual and audio streams, fused through an MLP and decoded via a Transformer under joint CTC/attention training. Experiments show that AUTO-AVSR achieves state-of-the-art performance with a word error rate (WER) of 0.9% on the LRS3 dataset, a 30% relative improvement over prior methods, while maintaining robustness to noise (WER $< 10\%$ at -7.5 dB SNR). Despite its superior accuracy, the framework's limitations include high model complexity (approximately 250 million parameters), significant computational cost, and performance saturation beyond about 1,500 hours of unlabelled data. Future work aims to enhance multi-lingual generalization and develop more efficient architectures for large-scale AVSR [3].

3. Conclusions

An AI-powered speech therapist assistant can bridge the accessibility gap in speech-language therapy by combining automatic speech recognition (ASR), pronunciation error detection, and NLP-driven feedback within a mobile-first, gamified platform. Leveraging pretrained ASR models such as Wav2Vec 2.0, Whisper, or QuartzNet fine-tuned on impaired speech datasets (e.g., CUChild, TinyMyST, MyST), the system can provide phoneme-level analysis and real-time corrective feedback through frameworks like Rasa or Dialogflow. Gamified exercises, storytelling, and AR features enhance engagement, while deployment on mobile apps ensures affordability and reach. Development can be achieved using Python for AI/ML, Hugging Face Transformers for speech and text processing, and Flutter or React Native for cross-platform interfaces, supported by cloud APIs (Google, Azure) where needed. The platform also includes a therapist dashboard that allows clinicians to monitor patient progress, view detailed speech analysis, and personalize therapy plans based on performance trends. Validation against expert SLP ratings using both objective (WER, phoneme accuracy, MOS) and subjective (satisfaction, progress) measures is essential to ensure accuracy and reliability. Future directions include personalizing therapy with reinforcement learning, adding multilingual support for under-represented languages, adopting privacy-first designs with secure or on-device processing, and integrating with telehealth platforms for hybrid clinician–AI therapy. This human-centered AI approach positions the tool not as a replacement but as an assistive, scalable companion for clinicians, caregivers, and patients.

References

1. A. Vinnarasu and V. D. Jose, "Speech to text conversion and summarization for effective understanding and documentation," *International Journal of Electrical and Computer Engineering (IJECE),* vol. 9, no. 5, pp. 3642–3648, *Institute of Advanced Engineering and Science,* October 2019. DOI:10.11591/ijece.v9i5.pp3642-3648.

 doi:10.20944/preprints202511.1793.v1

2. R. Jain, M. Y. Yiwere, D. Bigioi, P. Corcoran, and H. Cucu, "A Text-to-Speech Pipeline, Evaluation Methodology, and Initial Fine-Tuning Results for Child Speech Synthesis," *IEEE Access*, vol. 10, pp. 47628-47642, 2022. DOI: 10.1109/ACCESS.2022.3170836
3. Ma, P., Haliassos, A., Fernandez-Lopez, A., Chen, H., Petridis, S., & Pantic, M. (2023). "auto-avsr: audio-visual speech recognition with automatic labels," *arXiv preprint arXiv:2303.14307*.
4. Ali Babaeian, "Pronunciation Assessment: Traditional vs Modern Modes," *xitJournal of Education for Sustainable Innovation*, vol. 1, no. 1, pp. 61–68, 2023. DOI:10.56916/jesi.v1i1.530.
5. Suh, H., Dangol, A., Meadan-Kaplansky, H., Miller, C. A., & Kientz, J. A. (2024). "Opportunities and challenges for AI-based support for speech-language pathologists," *Proceedings of the 3rd Annual Symposium on Human-Computer Interaction for Work* (CHIWORK '24), 1–14. ACM.
6. A. Rahman, M. M. Kabir, M. F. Mridha, M. Alatiyyah, H. F. Alhasson, and S. S. Alharbi, "Arabic Speech Recognition: Advancement and Challenges," *IEEE Access, IEEE*, 2024.
7. S. B. Manir, K. M. S. Islam, P. Madiraju, and P. Deshpande, "LLM-Based Text Prediction and Question Answer Models for Aphasia Speech," *IEEE Access, IEEE*, August 2024.
8. J. Tobin, P. Nelson, B. MacDonald, R. Heywood, R. Cave, K. Seaver, A. Desjardins, P.-P. Jiang, and J. R. Green, "Automatic Speech Recognition of Conversational Speech in Individuals With Disordered Speech," *Journal of Speech, Language, and Hearing Research*, 2024.
9. Mariem Jelassi, Khouloud Matteli, Houssem Ben Khalfallah, and Jacques Demongeot, "Enhancing Personalized Mental Health Support Through Artificial Intelligence: Advances in Speech and Text Analysis Within Online Therapy Platforms," Information, vol. 15, no. 12, pp. 813, 2024. DOI: 10.3390/info15120813
10. A. Al-Ali, S. Al-Maadeed, M. Saleh, R. C. Naidu, Z. C. Alex, P. Ramachandran, R. Khodeeram, and R. K. M., "The Detection of Dysarthria Severity Levels Using AI Models: A Review," *IEEE Access*, 2024
11. R. Zeeshan, J. Bogue, and M. N. Asghar, "Relative Applicability of Diverse Automatic Speech Recognition Platforms for Transcription of Psychiatric Treatment Sessions," *IEEE Access*, Jul. 11, 2025.