SMART AI-ASSISTED SPOKEN LANGUAGE LEARNING APP FOR CHILDREN WITH SSD AND L2 LEARNERS

Mrs. E. Radhika[1] , Yamundla Rishi[2], Shivuni Vivek Sai[3], Vanamala Vishal[4] , Shaik Afrid[5] , Udurukota Rohith[6]

[1] *Assistant Professor, Department of CSE (ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING) TKR COLLEGE OF ENGINEERING & TECHNOLOGY*

[2,3,4,5] *UG Scholars in Department of CSE (ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING) TKR COLLEGE OF ENGINEERING & TECHNOLOGY*

ABSTRACT: The smart AI-Assisted Spoken Language Learning App is an innovative solution designed to support pronunciation learning and speech correction in children with Speech Sound Disorders and second language learners. Traditional speech therapy methods often face challenges related to accessibility, engagement, and scalability, especially in low-resource settings. This project leverages cutting-edge deep learning techniques to address these gaps by integrating Automatic Speech Recognition and pronunciation assessment into a gamified mobile application. The core engine is built on a multitask learning model using wav2vec2, which simultaneously performs speech recognition and evaluates pronunciation quality. The app encourages imitation-based learning through audio-visual prompts and motivational feedback, promoting high repetition, a key factor in effective speech intervention. Beyond child speech therapy, the system is designed to extend support to post-stroke and cardiac recovery patients with speech impairments. It incorporates Voice Activity Detection to isolate relevant speech segments and uses Integrated Gradients for model interpretability, ensuring transparent and explainable feedback. Real-time analytics and performance dashboards empower caregivers and speech-language pathologists to monitor progress and personalize therapy plans.

Keywords : SSD, ASR, wav2vec2, VAD, Pronunciation Scoring, Integrated Gradients, Fluency Dashboard, Speech Disfluency Classification.

1. Introduction

Speech and language development is essential for effective communication, academic success, and social interaction. However, children with Speech Sound Disorders (SSD) and second language learners often experience difficulties in achieving accurate pronunciation and fluency. Conventional speech therapy methods typically require frequent in-person sessions with trained professionals, making them expensive and inaccessible to many users. Furthermore, these methods often lack interactive elements, leading to reduced motivation among young learners. With advancements in Artificial Intelligence and mobile technologies, there is a significant opportunity to develop scalable and engaging solutions for speech learning. This paper presents an AI-assisted spoken language learning system that provides real-time feedback, personalized training, and gamified interaction to improve pronunciation skills. The proposed solution aims to bridge the gap between

traditional therapy and modern digital learning approaches by offering an accessible and adaptive platform.

A. Motivation:

Speech and language development during early childhood is critical for effective communication, academic success, and social integration. However, children with Speech Sound Disorders and second language -learners often face challenges in acquiring accurate pronunciation and fluency. Traditional speech therapy methods are resource-intensive, require frequent in-person sessions, and may lack engagement, especially for young learners. With the rise of Artificial Intelligence and mobile technologies, there is a growing opportunity to deliver scalable, personalized, and interactive speech therapy solutions that can bridge this accessibility gap. This project is motivated by the need to create a smart, AI-assisted speech therapy application that combines deep learning, gamification, and real-time feedback to support children with SSD and L2 learners. The goal is to empower caregivers, educators, and speech-language pathologists with a tool that enhances therapy outcomes while making the learning process enjoyable and adaptive.

2. Related Work

Recent advancements in speech processing and machine learning have significantly improved automated pronunciation assessment systems. Deep learning models such as wav2vec2 have demonstrated strong capabilities in extracting meaningful speech representations, enabling more accurate speech recognition and evaluation. Additionally, gamification techniques have been widely adopted in educational technologies to enhance user engagement and learning outcomes. Several studies have explored transformer-based models for speech disfluency detection and AI-based tools for pronunciation training. However, most existing systems lack explainability, making it difficult for users and therapists to understand how feedback is generated. Moreover, many applications are not optimized for child speech, which exhibits high variability and differs significantly from adult speech patterns. These limitations highlight the need for a more comprehensive system that combines accuracy, interpretability, and user-centric design.

3. System architecture

The system architecture of the proposed Smart AI-Assisted Spoken Language Learning Application is designed as a modular and scalable framework that integrates speech processing, deep learning, and user interaction components to deliver real-time pronunciation feedback. The architecture consists of multiple interconnected stages, beginning with audio acquisition and ending with feedback visualization. Initially, speech input is captured through a mobile or web-based user interface, where users are prompted to pronounce specific words or sentences. This raw audio input is then passed to the preprocessing module, which performs noise reduction, normalization, and silence removal to enhance signal quality and ensure that only relevant speech segments are processed. Voice Activity Detection (VAD) is employed at this stage to effectively isolate speech from background noise, improving the robustness and accuracy of subsequent analysis.

Following preprocessing, the cleaned audio signal is fed into the feature extraction module, where high-level speech representations are generated using a deep learning model based on wav2vec2. This model captures both phonetic and temporal characteristics of speech, enabling a comprehensive understanding of pronunciation patterns. The extracted features are then utilized by the speech evaluation module, which performs multitask learning to simultaneously carry out Automatic Speech Recognition (ASR) and pronunciation scoring. The system computes clarity and pronunciation scores by comparing the user's speech with predefined reference patterns, focusing on phoneme-level alignment and acoustic similarity. This dual evaluation mechanism ensures accurate assessment of both correctness and quality of speech production.

To enhance transparency and usability, the architecture incorporates an explainable AI component that provides interpretable feedback to users. Techniques such as Integrated Gradients are used to highlight specific parts of the speech signal that contribute to errors, enabling users to understand their mistakes and improve effectively. The feedback module then generates real-time responses in both visual and textual formats, guiding users toward correct pronunciation. Additionally, gamification elements such as scores, rewards, and progress indicators are integrated into the user interface to encourage repeated practice and sustained engagement.

The system also includes a backend infrastructure that manages data storage, user profiles, and performance tracking. Databases such as SQLite or Firebase are used to store session history, scores, and progress metrics, which are visualized through a dashboard accessible to caregivers and therapists. This dashboard enables continuous monitoring and analysis of user performance, facilitating personalized intervention strategies. Overall, the architecture ensures a seamless flow of data from input acquisition to feedback delivery, providing a real-time, adaptive, and user-friendly speech learning experience.

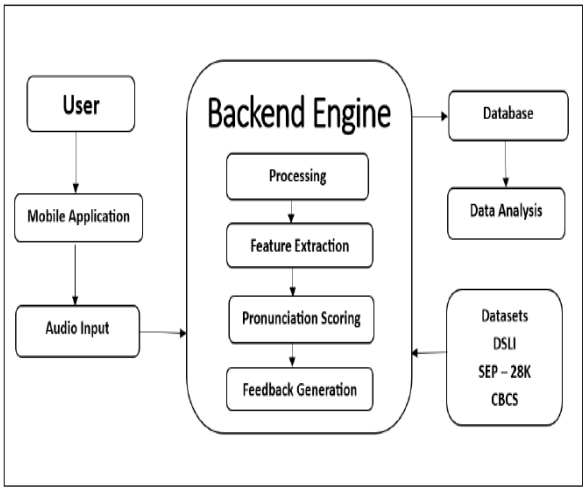

Fig:1. system architecture

4. Proposed System

The proposed system is a smart, AI-assisted speech therapy application designed to support children with Speech Sound Disorders, second language learners, and individuals recovering from speech impairments. It aims to bridge the gap between traditional therapy and modern digital solutions by offering a scalable, engaging, and clinically relevant platform for speech training. The system is built to function across mobile and web platforms, making it accessible for both home-based and clinical environments. At the core of the system lies a multitask deep learning

model based on wav2vec2, a state-of-the-art framework for self-supervised speech representation learning. This model is fine-tuned to simultaneously perform Automatic Speech Recognition and pronunciation scoring, enabling real-time feedback on speech quality.

The use of Voice Activity Detection ensures that only relevant speech segments are analyzed, improving accuracy and reducing noise interference. To enhance transparency and trust, the system integrates Integrated Gradients, an interpretability tool that helps explain how the model arrives at its decisions. The application adopts a gamified interface to make speech therapy more engaging for children. Users interact with visual and auditory prompts, perform imitation-based speech tasks, and receive feedback through intuitive elements like star ratings, progress meters, and animated cues. This approach encourages high repetition and sustained motivation—key factors in effective speech intervention. The feedback system is designed to be age-appropriate, using playful visuals and simplified scoring to help children understand and improve their pronunciation. To support caregivers and speech-language pathologists, the system includes a performance tracking dashboard that visualizes user progress over time. This dashboard provides insights into pronunciation accuracy, fluency trends, and therapy engagement, allowing for personalized therapy planning. The backend is built using Flask or FastAPI, with data stored securely in SQLite or Firebase, ensuring low-latency inference and robust data management. In Summary, the proposed system is a real-time, AI-powered speech therapy app tailored for children with SSD and L2 learners. It combines wav2vec2-based multitask learning, gamified feedback, and explainable AI to deliver engaging and adaptive therapy. Designed for both home and clinical use, it empowers caregivers and therapists with actionable insights and progress tracking.

5. Methodology

The system begins by capturing audio input through a mobile interface, where users are prompted to pronounce specific words or phrases. The recorded audio is preprocessed using noise reduction, normalization, and silence removal techniques to improve signal quality. The processed audio is then passed through a wav2vec2-based model, which extracts high-level speech features and generates phoneme-level representations. These features are used to compute clarity and pronunciation scores by comparing the user's speech with reference patterns. The system evaluates pronunciation accuracy based on phoneme alignment and confidence scores obtained from the ASR model. Additionally, prosodic features such as pitch and speech rate are analyzed to assess fluency. Based on these evaluations, the system generates personalized feedback in both textual and visual formats, guiding users toward correct pronunciation. The methodology ensures a seamless integration of speech processing, machine learning, and user interaction.

6. Implementation

The system is implemented using a modular architecture consisting of a frontend interface, backend server, and machine learning pipeline. The frontend is developed as a mobile or web application

that allows users to record speech, view feedback, and track progress. The backend is built using frameworks such as Flask or FastAPI, which handle API requests, audio processing, and communication with the machine learning model. Data is stored in databases like SQLite or Firebase to maintain user profiles, session history, and performance metrics. The system supports real-time inference, enabling immediate feedback after each speech attempt. Additional features such as user authentication, session tracking, and dashboard visualization are integrated to provide a comprehensive user experience.

7. Algorithms

A. Speech Representation & Feature Extraction

The core of the proposed system is a deep learning–based speech representation model that extracts meaningful features from raw audio signals. When a user records speech, the audio first undergoes preprocessing steps such as resampling to a fixed frequency (16 kHz) and silence removal using Voice Activity Detection (VAD). This ensures that only relevant speech segments are processed, improving model efficiency and accuracy. The processed audio is then passed through a pre-trained Wav2Vec2 model, which converts raw waveform inputs into high-dimensional feature embeddings. These embeddings capture phonetic structure, pronunciation patterns, and temporal speech characteristics. Unlike traditional handcrafted features (e.g., MFCC), Wav2Vec2 learns representations directly from raw audio, making it robust to variations in accent, tone, and speaking style. To derive meaningful scores, the extracted embeddings are aggregated and passed through a regression layer that predicts normalized speech quality. This output is transformed into a clarity score (0–100 scale). The model ensures consistent evaluation across different users and recording conditions, forming the foundation of the speech assessment pipeline.

B. Pronunciation Similarity & Imitation Scoring

For imitation-based tasks, the system evaluates how closely a user's speech matches a reference audio sample. Both the reference audio and user-recorded audio are processed through the same Wav2Vec2 pipeline to generate comparable embedding vectors. The system computes similarity using cosine similarity, which measures the angular distance between the two embedding vectors. A higher similarity score indicates closer alignment in pronunciation, rhythm, and articulation. The similarity value is then scaled and normalized into an imitation score (0–100). To improve robustness, embeddings are normalized before comparison, and scoring thresholds are applied to prevent inflated results. This method enables objective pronunciation assessment without requiring explicit phoneme-level alignment, making it efficient for real-time applications.

C. REST API Routing & Audio Processing Pipeline

The system follows a REST-based architecture to manage communication between the frontend, backend, and machine learning components. When a user completes a recording, the frontend sends a POST request containing the audio file and prompt ID to the backend server. Upon receiving the request, the backend performs

validation and preprocessing, including format checks, resampling, and speech segmentation. The processed audio is then passed through the Wav2Vec2 model for feature extraction and scoring. Based on the task type (normal or imitation), the system computes clarity and/or pronunciation scores and generates structured feedback. 36 The backend stores session data—including scores, timestamps, and prompt details—in the database. For data retrieval, the frontend sends GET requests to fetch dashboard metrics, session history, and performance trends. The backend responds with structured JSON data, which is dynamically rendered in the user interface. This modular pipeline ensures efficient, scalable, and real-time speech processing while maintaining clear separation between system components.

8. IMPLEMENTATION & RESULTS

A. Explanation of Key Functions

The AI-Based Speech Clarity Training System is built using a modular architecture combining frontend interaction, backend APIs, and deep learning–based speech processing. The system evaluates user speech, generates feedback, and tracks progress through multiple interconnected functions.

B. Speech Scoring Function

The core function of the system is speech evaluation implemented in the backend. When a user records audio, the system preprocesses it by resampling to 16 kHz and removing silence using Voice Activity Detection (VAD). The processed audio is passed through the Wav2Vec2 model to extract feature embeddings. These embeddings are fed into a regression layer to generate a normalized output, which is converted into a clarity score (0–100). This function ensures consistent and real-time speech quality evaluation.

C. Pronunciation (Imitation) Scoring Function

For imitation-based tasks, the system compares user speech with a reference audio sample. Both inputs are processed through the same feature extraction pipeline. Cosine similarity is computed between normalized embeddings to measure pronunciation similarity. The result is scaled into a 0–100 imitation score. This function enables objective evaluation of pronunciation, rhythm, and articulation.

D. Prompt Selection Function

The prompt selection function dynamically assigns practice prompts based on user level and previous activity. It ensures that prompts are not repeated unnecessarily and adapts difficulty using level mapping. This function supports structured and progressive learning.

E. Feedback Generation Function

Based on clarity score, imitation score, and speech duration, the system generates structured feedback. It categorizes performance into levels (low, medium, high) and provides actionable suggestions such as improving articulation, slowing speech, or matching reference audio. This function enhances learning through clear guidance.

F. User Progress & Gamification Function

The system tracks user performance using XP (experience points), levels, and streaks. After each session, XP is calculated and added to the user's total. Levels are updated dynamically, and streaks are computed based on daily activity. This function motivates consistent practice and engagement.

G. Guardian Insight Function

For child learners, the system generates analytical insights based on session history. It evaluates trends, stability, and risk levels using statistical measures such as mean and standard deviation. The function provides recommendations and early warnings to guide improvement.

H. Dashboard Aggregation Function

The dashboard function collects and processes user session data to compute metrics such as average scores, recent performance, improvement trends, and streaks. It prepares structured data for visualization in charts and progress indicators.

I. Session Storage Function

Each user session is stored with attributes such as timestamp, prompt ID, scores, and XP. This function ensures persistent tracking of user performance and supports analytics and history retrieval.

8.1 Method of Implementation

A. Frontend (Web Interface)

- Provides interactive UI for recording speech and viewing feedback
- Displays real-time states: recording, processing, results
- Includes dashboard with charts, streaks, and achievements
- Supports role-based experience (child learner / general user)

B. **Backend (FastAPI)**

- Handles API requests for scoring, prompts, and dashboard
- Processes audio and runs ML inference
- Manages authentication, sessions, and user data
- Generates feedback and analytics

C. Speech Processing Module (Deep Learning)

- Uses Wav2Vec2 for feature extraction
- Applies regression model for clarity scoring
- Uses cosine similarity for pronunciation evaluation
- Integrates VAD for speech segmentation

D. Database

- Stores user profiles, sessions, scores, and progress
- Maintains structured JSON-based storage (upgradeable to SQL/NoSQL)
- Supports retrieval for analytics and dashboard visualization

Deployment

- Web-based system accessible via browser
- Backend APIs serve real-time scoring and data
- Designed for scalability and future cloud deployment.

8.2 Forms

A. User Registration Form

The registration form collects Full Name, Email Address, and Password. The form applies client-side validation before submission. The name field requires a minimum of 3 characters containing only alphabetic characters. The email field accepts valid email formats. The password field enforces strong password rules requiring a minimum of 8 characters with at least one uppercase letter, one lowercase letter, one numeric digit, and one special character. Error messages are displayed inline below the form upon validation failure. Upon successful validation, the data is sent to the backend for secure account creation.

B. User Login Form

The login form collects Email Address and Password. Upon submission, the credentials are sent to the backend authentication API. If the login is successful, the user session is established using secure session management. The user is then redirected to the dashboard. If authentication fails, appropriate error messages are displayed to guide the user.

C. Profile Setup Form

The profile setup form allows users to select their role (e.g., learner or child learner) after initial login. If the child learner role is selected, additional fields such as guardian email may be collected. This role selection is fixed after submission and cannot be changed later, ensuring consistent user experience and personalized system behavior. The form data is stored in the backend and used for role-based feature activation such as guardian insights.

D. Speech Practice Interface

The speech practice interface allows users to interact with prompts and record their speech. It displays prompt text and, in imitation mode, provides reference audio playback. The user records audio using the microphone, and the system captures and sends the audio file to the backend via a POST request. A loading indicator is displayed during processing. Once scoring is complete, the system displays clarity score, pronunciation score, and structured feedback. Retry and next options are provided for continued interaction.

E. Dashboard Interaction (Analytics View)

The dashboard interface presents user performance data including total sessions, average scores, streaks, and progress trends. It dynamically fetches data from backend APIs and renders charts and metrics. Users can navigate to detailed history or achievements sections. The interface supports interactive elements such as time filters and audio-based summary playback. **F. Guardian Insight Interface (Child Mode)**

For child learners, an additional interface displays guardian insights based on performance analytics. It includes focus areas, recommendations, risk levels, and stability indicators. A text-to-speech option allows the insights to be read aloud. This interface is conditionally rendered based on user role and enhances monitoring and guidance.

8.3 OUTPUT SCREENS

The system produces the following output screens that enable user interaction, speech practice, performance tracking, and analytics visualization.

1. Login Screen

The login screen follows a modern UI design with a clean layout and centered authentication card. It displays a “Welcome Back” heading along with input fields for email and password. A login button initiates authentication, and error messages are displayed for invalid credentials. The interface ensures a simple and intuitive user entry point.

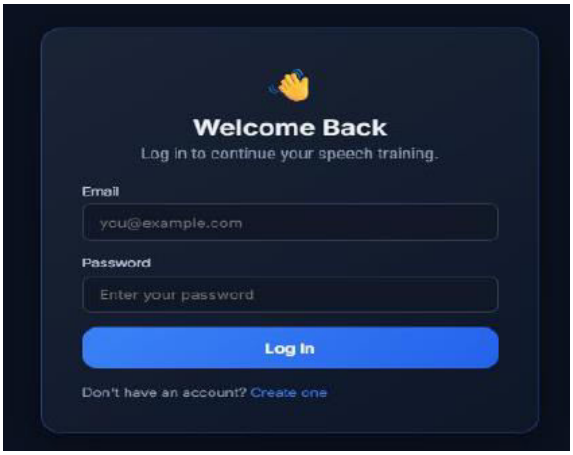

Fig: 8.1 Login Screen

2. Registration Screen

The registration screen allows new users to create an account by entering Full Name, Email, and Password. The interface includes inline validation for input fields, ensuring correct data entry before submission. Error messages are displayed dynamically to guide the user. The design maintains consistency with the login interface.

Fig : 8.2 Registration Screen

3. Profile Setup Screen

After login, users are directed to a profile setup screen where they select their role (e.g., learner or child learner). Additional fields such as guardian email may be displayed for child users. This screen ensures role-based customization of the system and restricts role modification after setup.

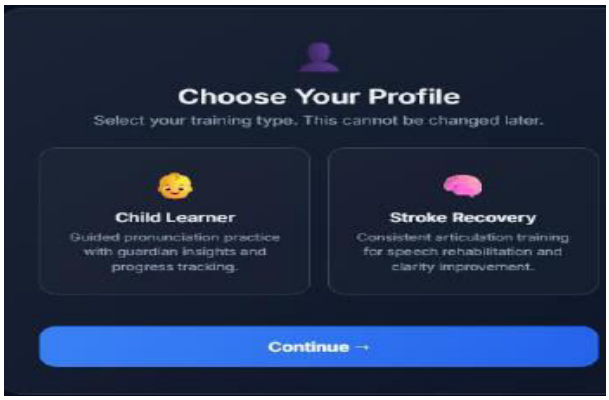

Fig : 8.3 Profile Setup Screen

4. Speech Practice Screen

The practice screen is the core interaction interface of the system. It displays the assigned prompt text and, in imitation mode, provides reference audio playback. Users can record their speech using the microphone, with real-time indicators such as recording status and audio visualization. After submission, the system processes the audio and displays feedback.

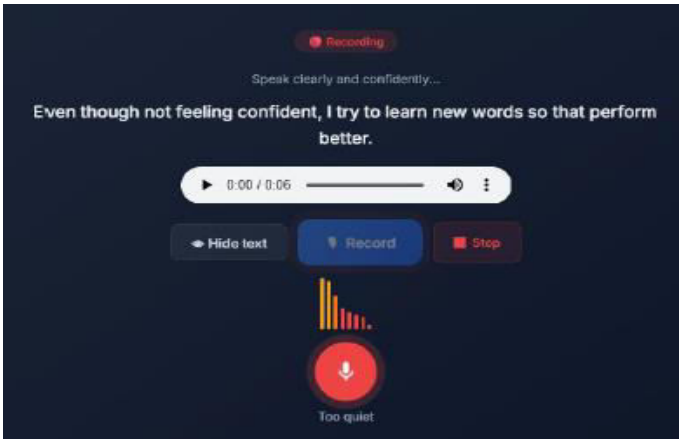

Fig : 8.4 Speech Practice Screen

5. Feedback & Results Screen

Once the speech is evaluated, the system displays a results panel showing clarity score, pronunciation score, and feedback messages. Visual elements such as progress bars and score indicators enhance readability. Options such as Retry and Next Prompt allow continued interaction and improvement.

Fig : 8.5 Feedback & Results Screen

6. Dashboard Screen

The dashboard provides a comprehensive overview of user performance. It displays metrics such as total sessions, average clarity, pronunciation scores, streak count, and current level. Graphs visualize performance trends over time, and progress bars indicate XP growth. The layout is structured to resemble a fitness-style performance tracker.

Fig : 8.6 Dashboard Screen

7. Guardian Insight Screen & Achievements

For child learners, an additional section displays guardian insights. It includes performance focus areas, recommendations, stability indicators, and risk levels. A text-to-speech option allows the insights to be read aloud. This screen supports monitoring and guided improvement. The system includes dedicated screens for viewing achievements and session history. Users can track past performance, review scores, and monitor improvements over time. The interface provides structured data presentation with clear navigation.

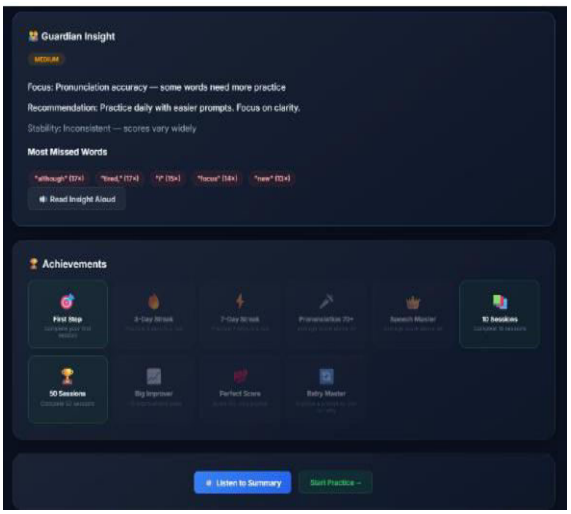

Fig : 8.7 Guardian Insight Screen & Achievements

8.4 Result Analysis

The result analysis evaluates the performance and effectiveness of the AI-Based Speech Clarity Training System across its key functional components, including speech evaluation, user interaction, and system responsiveness.

1. Speech Evaluation Performance

The system utilizes a Wav2Vec2-based pipeline for extracting speech features and generating clarity scores. The model demonstrates consistent performance across different users and recording conditions. Voice Activity Detection (VAD) effectively removes silence, ensuring that only relevant speech segments are evaluated. The clarity scoring mechanism produces stable outputs, enabling meaningful differentiation between low, moderate, and high-quality speech.

2. Pronunciation (Imitation) Scoring

The imitation scoring mechanism uses cosine similarity between user and reference embeddings. While the system successfully computes similarity in real time, observations indicate that embedding-based similarity can sometimes produce high scores even for imperfect matches due

to the generalized nature of learned representations. This highlights the need for further calibration or advanced alignment techniques for more precise pronunciation assessment.

3. Feedback Effectiveness

The feedback generation system provides structured and actionable insights based on user performance. Users receive guidance on clarity, pronunciation, and speaking pace, enabling iterative improvement. The inclusion of retry functionality supports reinforcement learning by allowing users to immediately apply feedback and observe improvements in subsequent attempts.

4. User Engagement & Gamification

The integration of XP, levels, streaks, and achievements significantly improves user engagement. The system successfully tracks progress over multiple sessions and provides visual motivation through dashboards and performance indicators. Users demonstrate improved consistency due to streak tracking and reward mechanisms.

5. Dashboard & Analytics Performance

The dashboard effectively aggregates session data and presents key metrics such as average scores, recent performance, and improvement trends. Graph-based visualization enables users to track their progress over time. The system maintains accurate computation of XP, levels, and streaks, supporting meaningful performance analysis.

6. Guardian Insight Evaluation

For child learners, the guardian insight module provides useful analytical summaries based on session history. It identifies trends, stability levels, and potential risks using statistical measures. The insights assist in monitoring progress and guiding improvement, although their effectiveness depends on sufficient session data.

7. System Performance

The system demonstrates efficient performance with a response time of approximately 2–4 seconds per request, including audio processing, model inference, and feedback generation. The FastAPI backend handles concurrent requests effectively, and the modular architecture ensures smooth interaction between frontend, backend, and ML components.

8. Overall System Effectiveness

The system successfully achieves its primary objectives of real-time speech evaluation, structured feedback delivery, and user progress tracking. It provides an interactive and scalable platform for speech training.

9. Conclusion

The present work successfully demonstrates the design and development of an intelligent AI-based speech clarity training system that integrates deep learning, real-time audio processing, and interactive user interfaces into a unified learning platform. The system provides an automated mechanism to evaluate speech clarity and pronunciation, delivering instant feedback and structured guidance to users through a seamless web-based experience.

The proposed system addresses a key limitation in traditional speech training approaches — the lack of accessible, consistent, and real-time feedback. By leveraging pretrained speech models for feature extraction and a trained scoring model for clarity assessment, the system evaluates user speech objectively. The integration of prompt-based practice,

imitation learning using reference embeddings, and adaptive feedback generation enables users to improve their speech through continuous, guided practice without the need for human supervision.

The system incorporates user profiling, progress tracking, and performance analytics through an interactive dashboard. Features such as session history, streak tracking, level progression, and achievement systems enhance user engagement and motivation. Additionally, the inclusion of a guardian insight module for child learners provides analytical summaries, risk indicators, and recommendations, extending the system's applicability to assisted learning scenarios.

The backend architecture built using FastAPI ensures efficient handling of audio processing, scoring, and API communication, while the frontend delivers a responsive and interactive user experience. The system demonstrates reliable performance with real-time processing capabilities and consistent output across multiple test scenarios. End-to-end integration of all modules — from audio capture to feedback and analytics — validates the robustness of the overall pipeline.

In conclusion, the project establishes that AI-driven speech analysis systems can significantly enhance language learning by providing scalable, real-time, and personalized training solutions. While certain components such as pronunciation similarity scoring require further refinement for production-level precision, the system provides a strong foundation for future enhancements including adaptive learning models, advanced phoneme-level analysis, mobile deployment, and large-scale educational integration.

10. REFERENCES

1. Hair, A., Ballard, K. J., Markoulli, C., Monroe, P., McKechnie, J., Ahmed, B., & Gutierrez-Osuna, R. (2021). A longitudinal evaluation of tablet-based child speech therapy with Apraxia World. ACM Transactions on Accessible Computing, 14(1), Article 3.

2. Wang, C., Rivière, M., Lee, A., Wu, A., Talnikar, C., Haziza, D., Williamson, M., Pino, J., & Dupoux, E. (2021). VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL), 993–1003.

3. Ibatova, G., Makhmetova, A., Zhoraeyeva, S. B., Amiresheva, B., Tinibekovna, N. S., & Satova, A. (2021). Psychological and pedagogical prerequisites of word-formation skills of Kazakh-speaking preschool children with speech disorder. World Journal on Educational Technology: Current Issues, 13(4), 1016–1028.

4. Cerezo, R., Calderón, V., & Romero, C. "A holographic mobile-based application for practicing pronunciation of basic English vocabulary for Spanish speaking children.", 2023.

5. Silva, T. F., Ribeiro, G. C. F., da Silva, C. E. E., Assis, M. F., Dezani, H., & Berti, L. C. "Efficacy in the use of gamification strategy in phonological therapy." CoDAS, vol. 35, no. 6, e20220181, 2023, pp. 1–8.

6. Georgiou, G. P. “Enhancing nonnative speech perception and production through an AI-powered application.”, 2023, pp.

7. Liu, J., Wumaier, A., Wei, D., & Guo, S. (2023). Automatic speech disfluency detection using wav2vec2.0 for different languages with variable lengths. Applied Sciences, 13(13), 7579. https://doi.org/10.3390/app13137579 .

8. Basak, K., Mishra, N., & Chang, H.-T. (2023). TranStutter: A convolution-free transformer-based deep learning method to classify stuttered speech using 2D Mel-spectrogram visualization and attention-based feature representation. Sensors, 23(19), 8033.

9. C. Deka, A. Shrivastava, and R. Kumar, “Towards Human-Centered AI in Speech Therapy: Perspectives from a Low-Resource Setting,” Research Square, Aug. 2024. doi: 10.21203/rs.3.rs-4833343/v1 .

10. Kim, Y., Kim, M., Kim, J., & Song, T.-J. (2024). Smartphone-based speech therapy for poststroke dysarthria: Pilot randomized controlled trial evaluating efficacy and feasibility. Journal of Medical Internet Research, 26, e56417.

11. A. Shrivastava and C. Deka, “Learning Pronunciation through Web-based Applications: An Exploratory Study,” International Journal of Advanced Science and Technology, vol. 29, no. 5, pp. 6531–6539, May 2024. 72

12. Shethia, U., Inamdar, V., & Kulkarni, V. (2025). Evaluating a digital speech therapy app for stuttering: A pilot validation study. arXiv preprint arXiv:2503.02743 [cs.HC].

13. Alhakbani, N., Alnashwan, R., Al-Nafjan, A., & Almudhi, A. (2025). Automated stuttering detection using deep learning techniques. Journal of Clinical Medicine, 14(10), 3552.

14. Kartika, S. “AI and Gamification: Exploring the Integration of Generative AI in Interactive English Language Learning Apps.” International Journal of Language, Humanities, and Education (IJLHE), vol. 8, no. 1, pp. 105–114, June 2025.

15. Fatima, N., & Faraz, H. “Accelerating Second Language Learning through Artificial Intelligence: A Study of AI-Driven Personalized Learning Platform.”, March 2025.

16. Georgiou, G. P. “Enhancing Nonnative Speech Perception and Production through an AI-Powered Application.” -2025.

17. Nguyen, T., Fredouille, C., Ghio, A., Balaguer, M., & Woisard, V. “Exploring ASR-Based Wav2Vec2 for Automated Speech Disorder Assessment: Insights and Analysis- Oct 2025.

18. Getman, Y., Phan, N., Al-Ghezi, R., Voskoboinik, E., Singh, M., Grósz, T., Kurimo, M., Salvi, G., Svendsen, T., Strömbergsson, S., Smolander, A., & Ylinen, S. “Developing an AI-Assisted Low-Resource Spoken Language Learning App for Children” – Aug 2023.