Est. 1841 | YORK ST JOHN UNIVERSITY

Samarasekara, Thilina Lakshan, Rubasin Siriwardana, Samanthi, Weerasinghe, Lokesha, Weerasinghe, Thulasika Nayananjalee, Wijesekara, Thinama Renugi and Liyanage, Himakara (2026) Artificial intelligence-driven Sinhala auditory–verbal tutoring system for hearing-impaired children. International Journal of Systematic Innovation.

Downloaded from: https://ray.yorksj.ac.uk/id/eprint/16391/

The version presented here may differ from the published version or version of record. If you intend to cite from the work you are advised to consult the publisher's version: https://doi.org/10.6977/IJoSI.202608_10(4).0006

Research at York St John (RaY) is an institutional repository. It supports the principles of open access by making the research outputs of the University available in digital form. Copyright of the items stored in RaY reside with the authors and/or other copyright owners. Users may access full text items free of charge, and may download a copy for private study or non-commercial research. For further reuse terms, see licence terms governing individual outputs. Institutional Repositories Policy Statement

RaY

Research at the University of York St John

For more information please contact RaY at ray@yorksj.ac.uk

ACCSCIENCE PUBLISHING

International Journal of Systematic Innovation

ARTICLE

Artificial intelligence-driven Sinhala auditory–verbal tutoring system for hearing-impaired children

Thilina Lakshan Samarasekara[1], Samanthi Rubasin Siriwardana[2], Lokesha Weerasinghe[1]*, Thulasika Nayananjalee Weerasinghe[1], Thinama Renugi Wijesekara[1], and Himakara Liyanage[1]

[1]Department of Information Technology, Faculty of Computing, Sri Lanka Institute of Information Technology, Malabe, Western Province, Sri Lanka

[2]Department of Computer and Data Science, York St John University, London, United Kingdom

Abstract

Speech therapy for hearing-impaired children in Sri Lanka is significantly limited by factors such as cost constraints and the lack of automated, language-specific tools to aid in therapy sessions. This study presents a Sinhala auditory training platform for hearing-impaired children that combines adaptive learning with automatic pronunciation evaluation. The platform uses Bayesian knowledge tracing and multi-armed bandit task sequencing to estimate learner competence and personalize training. Its task system is based on four language comprehension task types that serve as blueprints for automatically generating activities of varying difficulty, supported by 353 audio assets. The platform covers phoneme discrimination, syllable processing, and word recognition, while providing analytics for therapists. A pronunciation evaluation module built on a fine-tuned Wav2Vec2 model delivers phoneme-level feedback using forced alignment and goodness of pronunciation scoring. To support this module, 1,534 Sinhala speech recordings from hearing-impaired speakers were collected for training. Experimental results using the model achieved a phoneme error rate of 0.289, demonstrating stable convergence during training. The system highlights the feasibility of combining adaptive tutoring and speech-based feedback for scalable, personalized auditory rehabilitation in Sinhala.

Keywords: Auditory–verbal therapy; Hearing impairment; Sinhala speech therapy; Wav2Vec2; Bayesian knowledge tracing

***Corresponding author:**
Lokesha Weerasinghe
(Lokesha.w@sliit.lk)

Citation: Samarasekara, T. L., Siriwardana, S. R., Weerasinghe, L., Weerasinghe, T. N., Wijesekara, T. R. & Liyanage, H. (2026). Artificial intelligence-driven Sinhala auditory–verbal tutoring system for hearing-impaired children. *Int J Systematic Innovation*, 10(4): 026140043. https://doi.org/10.6977/IJoSI.202608_10(4).0006

Received: April 2, 2026

Revised: July 9, 2026

Accepted: July 22, 2026

Published online: August 21, 2026

Copyright: © 2026 Author(s). This is an Open-Access article distributed under the terms of the Creative Commons Attribution License, permitting distribution, and reproduction in any medium, provided the original work is properly cited.

Publisher's Note: AccScience Publishing remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

Hearing loss affects many children worldwide, with a significant number born with permanent impairment. In Sri Lanka, school-age children with hearing loss highlight the need for accessible rehabilitation. Cochlear implants improve auditory access by converting sound into electrical signals for the auditory nerve, but children require structured rehabilitation to develop listening, phonological, and spoken language skills after the process (Yong-Xin *et al.*, 2007).

Auditory–verbal therapy (AVT) is a common rehabilitation approach that helps hearing-impaired children develop listening and spoken-language skills through guided

practice across detection, discrimination, identification, and comprehension. However, its traditional therapist-led and manually assessed form can limit therapy intensity, continuity, and accessibility (Colibaba *et al*., 2022; Fernando *et al*., 2024).

For this reason, computer-based auditory training has become an important complementary approach. It enables repeated, self-directed practice beyond face-to-face sessions and can help increase training frequency in home environments (Gnadlinger *et al*., 2024; Ishihara & Tsuda, 2021), giving the person extra time to improve. This makes adaptive learning especially valuable, as it allows a system to estimate learner competence and personalize activities in ways that better support steady progress (Gnadlinger *et al*., 2024; Segal *et al*., 2018; Seo, 2017; van de Sande, 2013).

Despite advances in digital rehabilitation technologies, low-resource languages such as Sinhala remain largely underserved. Existing applications developed for other linguistic regions are not fully suitable for this setting, as auditory and phoneme-level practice must align with the learner's mother tongue, especially during early spoken-language rehabilitation (Lee & Tsoi, 2021; Liu & Fu, 2007; Nonis, 2015). In the Sri Lankan context, this creates an important gap: hearing-impaired children have highly limited access to Sinhala-supported digital platforms that can provide adaptive listening and pronunciation practice matched to local language needs (Fernando *et al*., 2024). With this gap identified, this work focuses on designing a platform that supports broader language development while prioritizing Sinhala speech practice in a form more suitable for local rehabilitation use.

The proposed platform targets hearing-impaired children aged 5–13 and integrates an adaptive auditory training framework that personalizes activities using Bayesian knowledge tracing and multi-armed bandit task selection (Segal *et al*., 2018; van de Sande, 2013), together with a fine-tuned Wav2Vec2.0-based pronunciation evaluation module that analyzes spoken responses and provides phoneme-level feedback through forced alignment and goodness of pronunciation metrics (Witt & Young, 2000). These intelligent tutoring mechanisms continuously analyze learner performance to estimate mastery and select suitable activities, helping maintain an effective balance between challenge and skill level to improve engagement and skill acquisition.

2. Related work

Several digital auditory rehabilitation systems have been developed to support individuals using cochlear implants and hearing aids. Platforms such as Angel Sound, AB CLIX, LACE (Listening and Communication Enhancement), and AudioCardio demonstrate how computer-based auditory training can deliver structured listening exercises, guided practice, and performance tracking to improve speech perception and listening abilities (Gnadlinger *et al*., 2024; Ishihara & Tsuda, 2021).

Most existing platforms are designed primarily for English or other widely used languages. During early spoken-language development, auditory training must closely align with the learner's mother tongue, because phonological structures such as phoneme inventories, syllable patterns, and contrastive sound features differ significantly across languages (Lee & Tsoi, 2021; Liu & Fu, 2007; Nonis, 2015). For Sinhala-speaking children, effective auditory training therefore requires exercises built around Sinhala phonemes, vowel-length contrasts, culturally familiar vocabulary, and locally recorded speech materials (Nonis, 2015; Weerasinghe *et al*., 2005). While existing platforms provide valuable architectural frameworks for digital auditory training, there remains a clear need for a localized auditory rehabilitation system tailored to the phonological structure and linguistic characteristics of the Sinhala language (Fernando *et al*., 2024; Nonis, 2015; Weerasinghe *et al*., 2005).

As shown in Table 1, a high contrast can be observed between the existing and proposed systems.

Table 1. Comparison across systems

System	Target users	Language/Limitation	Relevance to our work
Angel Sound	CI/HA users, mainly adults/ teenagers	Built for major languages, not Sinhala	Shows value of structured digital listening practice
AB CLIX	Adult CI users	English-based word contrasts; adult-oriented	Supports graded discrimination training design
LACE	Adult hearing-aid users	English sentence materials; limited child suitability	Demonstrates adaptive speech-in-noise training
Proposed system	Hearing-impaired children	Localized for Sinhala and child learning needs	Addresses language, age, and content gap

Abbreviations: CI: Cochlear implant; HA: Hearing aid.

3. System overview

The platform targets hearing-impaired children (aged 5–13) and supports therapist-guided and home-based practices. The system's core use cases include children completing short and well-structured training sessions, adaptive selection of subsequent personalized tasks, and therapist access to dashboards for monitoring mastery and trends. The curriculum is implemented as modular activities, including phoneme-level training (vowel/consonant identification and contrasts), sound discrimination tasks, syllable-count tasks supported by syllabification rules, and controlled word or sentence-length tasks with adjustable difficulty (Nonis, 2015; Weerasinghe *et al.*, 2005). The merged system integrates an additional module for pronunciation evaluation. For tasks that require spoken responses, the pronunciation engine performs rapid validation and produces interpretable confidence scores and customized feedback (Witt & Young, 2000; Yang *et al.*, 2022).

The proposed system was implemented using a FastAPI backend for handling API requests, a React frontend for interactive user interfaces, and PostgreSQL for data storage. Figure 1 summarizes the functionality of the overall system architecture, highlighting the operational logic.

4. Methodology

An evidence-centered design approach was adopted to infer auditory and phonological competence from observable learner behaviors, such as response accuracy, response latency, and hint usage. Task design was guided by therapist-based auditory training manuals and adapted to the phonological structure of Sinhala. All these findings provided a clinically grounded and linguistically appropriate basis for the methodology.

4.1. Dataset and linguistic foundation

The hearing-training dataset was not treated as a fixed exercise bank but as a structured instructional source from which task-ready auditory units were systematically derived. Clinician-guided cochlear implant rehabilitation manuals were analyzed to extract core training patterns, including Ling sound practice (/a/, /e/, /i/, /u/, /ʃ/, /s/, /m/), environmental sound recognition with paired visual cues, syllable-count discrimination, and vowel–consonant contrast tasks at both word and sentence level (Colibaba *et al.*, 2022; Fernando *et al.*, 2024; Gnadlinger *et al.*, 2024). Sentence materials were organized by length, while word and sentence pairs from the manuals were decomposed and tokenized into similarity-based clusters to support progression from easier to more difficult levels. This conversion allowed clinically grounded content to be reshaped into a level-aware dataset suitable for adaptive task generation rather than simple static playback.

One of the major challenges in training pronunciation assessment models is the scarcity of labeled phoneme data, especially for second-language or impaired speech datasets (Yang *et al.*, 2022). To create a pronunciation model dataset, speech data were collected from five hearing-impaired children undergoing speech therapy sessions at

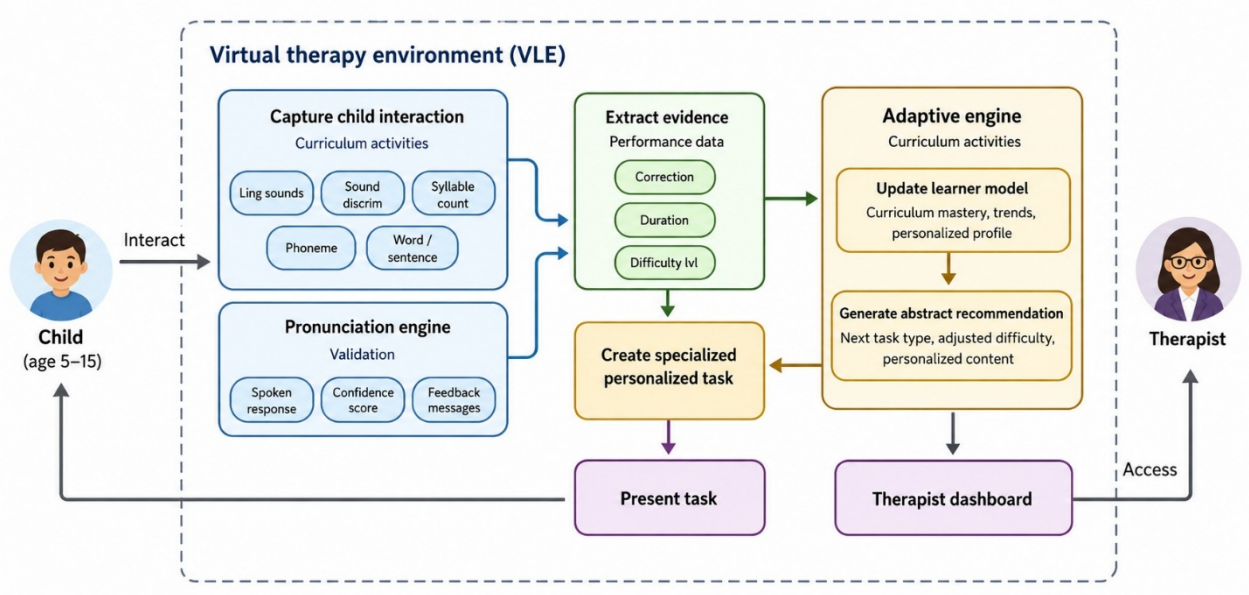


Figure 1. Overall system architecture

clinical therapy centers. The participants were between 8 and 11 years of age and were selected with the assistance of therapists during regular therapy visits. Prior to the recordings, informed consent forms were obtained from the children and their parents, allowing the recordings to be used strictly for research purposes while ensuring participant privacy and ethical compliance. Overall, the dataset represents a balanced mix of therapy-attending children within the selected age group.

Each participant took part in a recording session of approximately 15 minutes. During the session, participants were instructed to pronounce each target word 20 consecutive times to capture pronunciation variability and produce a robust dataset for analysis. Recordings were conducted in a quiet indoor environment using a mobile phone to minimize background noise and ensure acceptable audio quality.

The raw audio recordings were subsequently segmented into 1,534 meaningful utterance-level chunks using Audacity version 3.7.7, ensuring that each segment corresponded to a single target word without background interference or overlapping speech.

For consistency and compatibility with the Wav2Vec 2.0 feature extractor (Baevski *et al.*, 2020), all audio samples were standardized through a normalization process. Specifically, each recording was converted to a 16 kHz sampling rate, mono-channel configuration, and 16-bit WAV format. This preprocessing step ensures uniform input representation and optimal performance during feature extraction and model training (Yang *et al.*, 2022).

Table 2 summarizes the characteristics of the speech dataset used for model development and evaluation.

4.2. Word selection and phonetic categories

The target word list was adapted from a previous study (Nonis, 2015), which documented early phonetic acquisition patterns in Sinhala-speaking children. From this source, 20 words were selected, covering five major phonetic groups in Sinhala:

- Stops are produced by completely blocking the airflow in the vocal tract and then releasing it suddenly: balla (W1), ibba (W2), dath (W3), poddak (W4)
- Nasal consonants are produced when the airflow passes through the nasal cavity: samanalaya (W5), maduruwa (W6), lunu (W7), annasi (W8)
- Approximants involve articulators approaching each other without creating significant friction in the airflow: hawa (W9), yathura (W10), tharuwa (W11), oluwa (W12)
- Fricatives are produced by forcing air through a narrow constriction in the vocal tract, creating turbulent airflow and audible friction: ahasa (W13), gasa (W14), katha (W15), saban (W16)
- Clusters involve sequences of two or more consonants without intervening vowels, requiring complex articulatory coordination: aiskrim (W17), sapaththuwa (W18), prashna (W19), kridawa (W20)

These categories were chosen to represent diverse articulatory and phonological patterns (Nonis, 2015), enabling the pronunciation model to capture both vowel variation and consonant complexity, particularly the increased difficulty associated with consonant clusters in hearing-impaired speech (Lee & Tsoi, 2021; Nonis, 2015).

Table 2. Characteristics of the speech dataset used for model development and evaluation

Item	Value
No. of participants	5
Age (years)	8–11
Total recordings	1,534
Words	20
Recordings per word	≈76
Samples used in evaluation	993

5. System implementation and architecture

5.1. Target users and use cases

The primary users of the platform were hearing-impaired children between the ages of 5 and 13 undergoing AVT training. The main use cases supported by the platform are as follows:

- The learner completes short auditory training activities, such as sound identification, phoneme discrimination, syllable counting, and word-length recognition.
- The adaptive engine analyzes performance and selects the next training activity based on the learner's estimated mastery level (Segal *et al.*, 2018; van de Sande, 2013).
- Therapists and caregivers review learner progress through visual dashboards that summarize mastery trends, engagement levels, and error patterns. This structure allows the system to provide personalized practice (Fernando *et al.*, 2024; Gnadlinger *et al.*, 2024).

Although the platform is designed for children aged 5–13 years, the pronunciation evaluation model was trained using recordings collected only from children aged

8–11 years. Consequently, the reported speech recognition results cannot yet be generalized to younger users, whose articulation patterns may differ substantially. The adaptive tutoring components remain applicable across the broader target age range because they are independent of the acoustic model, whereas future work will expand the speech corpus to include children aged 5–7 and 12–13 years.

5.2. Modules and task types

The curriculum was implemented as different modular activities:

- Environmental sound identification: identification of common everyday sounds, such as fridge humming, kettle boiling, man's voice, air conditioning, car engine from inside the car, wind, fan, and a zipper sound (Colibaba *et al*., 2022; Ishihara & Tsuda, 2021).
- Phoneme-level training: vowel and consonant identification and contrasts, including short–long vowel distinctions (Lee & Tsoi, 2021; Liu & Fu, 2007; Nonis, 2015).
- Ling sound training: /m/, /u/, /a/, /i/, /ʃ/, /s/ detection/ discrimination (if included in the platform scope).
- Syllable-count tasks: rule-based syllabification → syllable number selection (Weerasinghe *et al*., 2005).
- Word/sentence length tasks: controlled increases in length/difficulty with noise/variability options (Fernando *et al*., 2024; Gnadlinger *et al*., 2024).

5.3. Multi-layer adaptive architecture

5.3.1. Layer 1: Learner state estimation (Bayesian knowledge tracing mastery per skill)

Each skill was defined at the phoneme or phoneme-class level (e.g., Sinhala vowels, vowel-length contrasts, consonant classes). After each item response, the system updates mastery probability *P*(K) using Bayesian knowledge tracing parameters (initial mastery, learning rate, slip, guess), as introduced by Corbett and Anderson (1994) and described in subsequent analyses (Liu & Fu, 2007). The mastery state supports:

- remediation targeting for low *P*(K),
- consolidation for near-threshold *P*(K),
- advancement when sustained mastery is observed.

$$P(K_n|Correct) = \left(P(K_{n-1})(1-S)\right)\left(P(K_{n-1})(1-S)+\left(1-P(K_{n-1})\right)G\right) \quad (1)$$

$$P(K_n|Incorrect) = \left(P(K_{n-1})S\right)/\left(P(K_{n-1})S+\left(1-P(K_{n-1})\right)(1-G)\right) \quad (2)$$

where $P(K_{n-1})$ is the prior mastery probability before the current item, S is the slip probability, G is the guess probability, T is the learning transition probability, $P(K_n|Obs)$ is the posterior mastery probability after observing the learner response, and $P(K_n^*)$ is the updated mastery probability after applying the learning transition.

5.3.2. Layer 2: Candidate item generation (constraints and content safety)

A candidate set was built using:

- prerequisite constraints (e.g., phoneme discrimination before syllable counting),
- age/level constraints,
- avoidance rules (e.g., limit repeated exposure, prevent fatigue).

This layer prevents purely reward-driven sequencing from recommending tasks that require missing prerequisites, aligning with computerized adaptive testing (CAT)-style constraint awareness and content balancing principles (Seo, 2017).

5.3.3. Layer 3: Next-item selection policy (bandits + difficulty control)

The next-item selector was modeled as a policy that chooses among candidate arms (e.g., phoneme × item type × difficulty). The system supports:

- Thompson sampling (Bayesian exploration),
- Upper confidence bound (UCB; optimistic exploration via confidence bounds),
- LinUCB (contextual selection using features such as mastery gap, response time, noise level, recent error streak).

Bandit personalization was justified by Multi-Armed Bandits-based Personalized Learning Environment (MAPLE)'s formulation of sequencing as maximizing expected learning gains over time (Seo, 2017). Reward signals were defined using measurable learning proxies, such as:

- correctness improvement trend,
- decreased response time for a skill,
- reduced hint usage,
- mastery probability increase Δ*P*(K).

$$R_t = \alpha\Delta Acc_t + \beta\Delta Speed_t + \gamma\Delta Hint_t + \delta\Delta P(K)_t \quad (3)$$

$$\Delta Acc_t = Acc_t - Acc_{t-1} \quad (4)$$

$$\Delta Speed_t = \frac{RT_{t-1} - RT_t}{RT_{t-1}} \quad (5)$$

$$\Delta Hint_t = \frac{Hint_{t-1} - Hint_t}{Hint_{t-1}} \quad (6)$$

where:

$$\Delta P(K)_t = P(K)_t - P(K)_{t-1} \quad (7)$$

5.3.4. Layer 4: Scaffolding and prompting (hinting with fade out)

The platform provides multiple levels of support to assist learners, including replaying audio, reducing the number of distractor options, offering visual cues, and giving hints of increasing difficulty. Scaffolding was framed as temporary support that fades as competence increases (Yang *et al*., 2022). Prompting strategies were mapped to established prompting hierarchies studied in applied behavior analysis, supporting structured independence-building rather than ad hoc feedback.

5.3.5. Layer 5: Monitoring, data quality, and safety analytics

To prevent misinterpretation of incorrect responses caused by fatigue, random tapping, or disengagement, the system applies monitoring heuristics:

- rolling accuracy and moving averages,
- response-time anomalies,
- streak-based disengagement flags,
- optional cumulative sum (CUSUM)-based aberrance detection (dual thresholds) motivated by work applying CUSUM to detect aberrant responses.

Where implemented, person-fit and psychometric misfit concepts motivate isolating unreliable segments to improve score validity and the reliability of mastery estimates.

As illustrated in Figure 2, it is clear how each layer exists in the multi-layer architecture.

Table 3 summarizes all the modules and their usage purposes.

Bayesian knowledge tracing was preferred over deep knowledge tracing approaches because it provides interpretable mastery estimates, requires substantially fewer learner interactions, and is well suited to small educational datasets typical of clinical rehabilitation settings. Multi-armed bandit algorithms were selected because they efficiently address sequential task selection by balancing exploration and exploitation while remaining computationally lightweight for real-time deployment. Together, these methods provide a transparent and data-efficient adaptive tutoring strategy suitable for low-resource auditory rehabilitation.

5.4. Phoneme inventory and learning targets

The curriculum adopted Sinhala phoneme-level targets consistent with descriptions of Sinhala segmental phonology (40 phonemes: 14 vowels, 26 consonants). This supports the following requirements:

- per-phoneme mastery tracking,
- explicit vowel-length contrast training,
- structured progression from high-frequency phonemes to expanded coverage.

5.5. Orthography-aware design: Abugida constraints

Sinhala is an abugida in which consonant symbols typically encode an inherent vowel, and vowel signs modify the consonant form. Therefore, the system prioritizes phonetic targets (what is heard) over grapheme naming (what is written). This reduces the risk of conflating orthographic conventions with auditory categories and supports clearer skill definitions for mastery modeling (Nonis, 2015; Weerasinghe *et al*., 2005).

5.6. Model implementation

The pronunciation evaluation component was implemented using a deep speech recognition model based on the Wav2Vec2 architecture (Baevski *et al*., 2020; Yang *et al*., 2022).

A custom phoneme vocabulary containing 21 Sinhala phonemic tokens (Yang *et al*., 2022) was constructed. The tokens represent common phoneme units used in the pronunciation training dataset. A connectionist temporal classification (CTC)-based tokenizer (Graves *et al*., 2006) and processor were implemented using the Wav2Vec2 framework. Special tokens such as padding and unknown tokens were included to satisfy the requirements of the CTC training objective.

The pretrained Facebook/wav2vec2-xls-r-300m (Babu *et al*., 2021; Morris *et al*., 2004) model was used as the base acoustic model. Several configuration parameters were adjusted to match the target vocabulary and ensure stable training:

- CTC loss reduction set to mean,
- zero-infinity loss handling enabled,
- vocabulary size adjusted to match the phoneme token set.

To adapt the pretrained model to Sinhala speech produced by hearing-impaired speakers, a two-stage fine-tuning strategy was employed.

In the first phase (epochs 0–5), the feature encoder remained frozen, ensuring that the pre-trained acoustic representations remained stable while the randomly initialized CTC (Graves *et al*., 2006) head learned the mapping between extracted audio features and Sinhala phonemes. This stabilization period allowed the classification layer to adapt without disrupting the underlying learned representations.

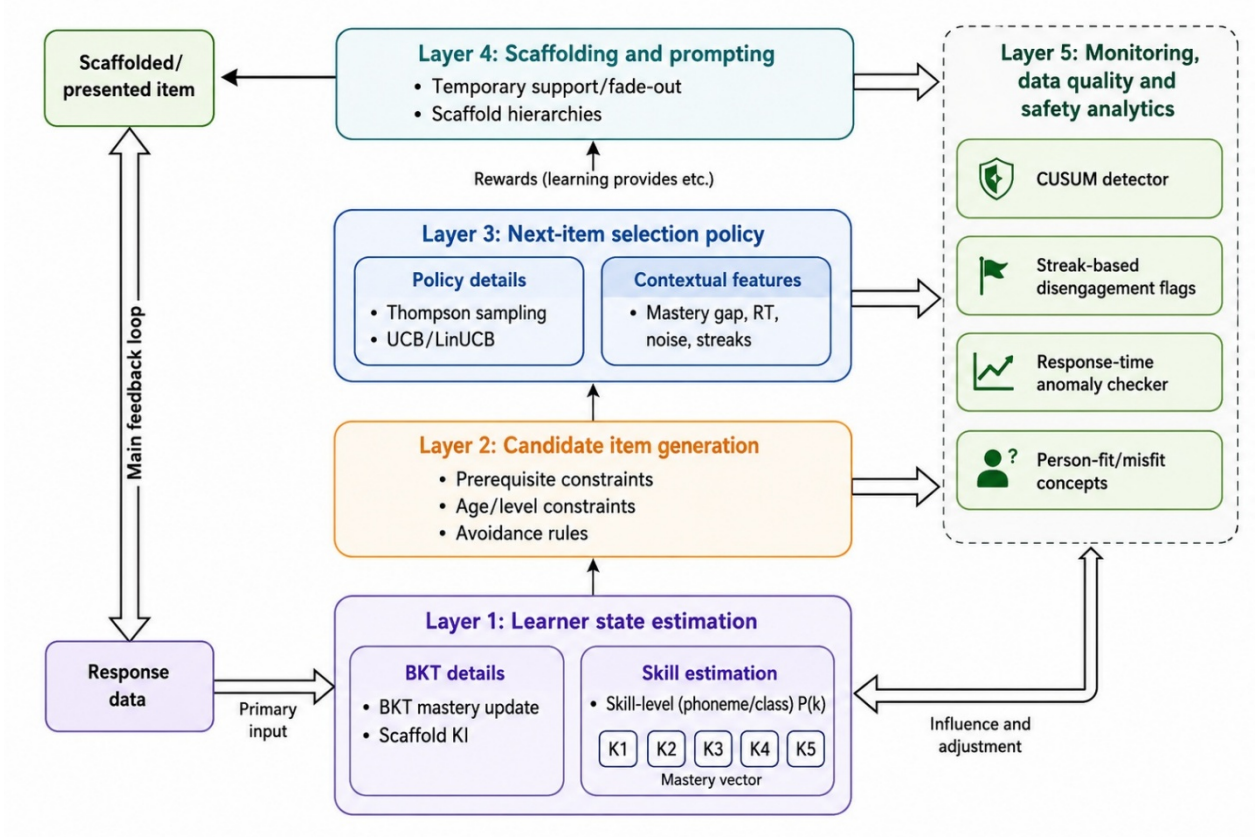


Figure 2. Multi-layer adaptive architecture
Abbreviations: BKT: Bayesian knowledge tracing; KI: Knowledge index; RT: Reaction time; UCB: Upper confidence bound.

Table 3. Module summarization

Module	Purpose	Logic	Key values
M0	Profiling + scaffolding	Rubric scoring, weakest-domain targeting, adaptive scaffolding	L1: <0.25; L2: 0.25–0.49; L3: 0.50–0.74; L4: ≥0.75; mastery: 0.80; concern: 0.30; regression: −0.10; min obs.: 5; stability: 3; transition: ≥0.70 (overall), ≥4/7 (L4), scaffold: ≤1.5
M1	Adaptive assessment	CAT-style difficulty control	Upper: 0.70; lower: 0.40; window: 5; level: 1–3
M2	Learning recommendation	Knowledge gap + prerequisite ordering + weighted ranking	Gap: 0.50; mastery: 0.80; engagement: 0.30/0.70; completion: 0.90; style conf.: 0.60; min interactions: 10
M3	Real-time intervention	Signal monitoring + weighted trigger + hint progression	Trigger: 0.60; errors: ≥3; stagnation: >2*x*; rapid: <5 s; inactivity: >60 s; frustration: >0.70; hint levels: 3
M4	Prediction + risk	Weighted prediction + risk tiers + action rules	High: <0.40; medium: <0.60; pass: 0.50; excellence: 0.85; trend: 5; min data: 3; improvement: 0.05

Abbreviations: CAT: Computerized adaptive testing; conf.: Confidence; obs.: Observations.

In the second phase (epochs 6–30), the feature encoder was unfrozen to allow end-to-end fine-tuning. This approach enabled the model to refine its acoustic filters to better capture characteristics of hearing-impaired speech while maintaining stable pretrained representations.

Training was conducted using a Google Colab GPU environment equipped with an NVIDIA T4 accelerator, which provided Compute Unified Device Architecture (CUDA)-enabled hardware support for efficient model fine-tuning. Due to the large parameter size of the wav2vec2-xls-r-300m model and the memory limitations of the available GPU, the training process was performed using a batch size of 4 samples per iteration.

The model was trained using the AdamW optimizer with a learning rate of 1×10^{-5}, and Gradient clipping with a maximum norm of 1.0 was applied to prevent unstable parameter updates. Training convergence was monitored using the CTC loss metric. Training was performed for 30 epochs. Figure 3 illustrates only the first 10 epochs because most convergence occurred during this period. The reported final CTC loss (1.1343) corresponds to epoch 9, after which only marginal improvements were observed, indicating that end-to-end fine-tuning allowed the model to adapt its acoustic representations more effectively to the characteristics of hearing-impaired speech (Yang *et al.*, 2022).

5.7. Pronunciation evaluation

During training sessions, learner speech was analyzed using a multi-stage pronunciation evaluation pipeline. First, the acoustic model generates frame-level emission probabilities from the input speech signal. These emissions are decoded into phoneme sequences using a greedy CTC decoding (Graves *et al.*, 2006) procedure.

Next, a similarity validation stage compares the predicted phoneme sequence with the expected phoneme sequence of the target word. The similarity score is calculated using the Levenshtein distance metric. Utterances with extremely low similarity scores are rejected, and the learner is prompted to retry. For valid utterances, the system performs forced phoneme alignment using torchaudio alignment functions. The alignment process synchronizes predicted phonemes with time frames of the audio signal.

Finally, pronunciation quality is estimated using goodness of pronunciation scores (Witt & Young, 2000) derived from phoneme-level log probabilities. The scores are normalized using min–max scaling to produce a final confidence value between 0 and 1. This confidence score is translated into user-friendly feedback messages that guide learners toward improved articulation (Witt & Young, 2000).

6. Results and discussion

This section presents the experimental evaluation of the proposed system, focusing on both pronunciation recognition performance and the behavior of the adaptive training framework.

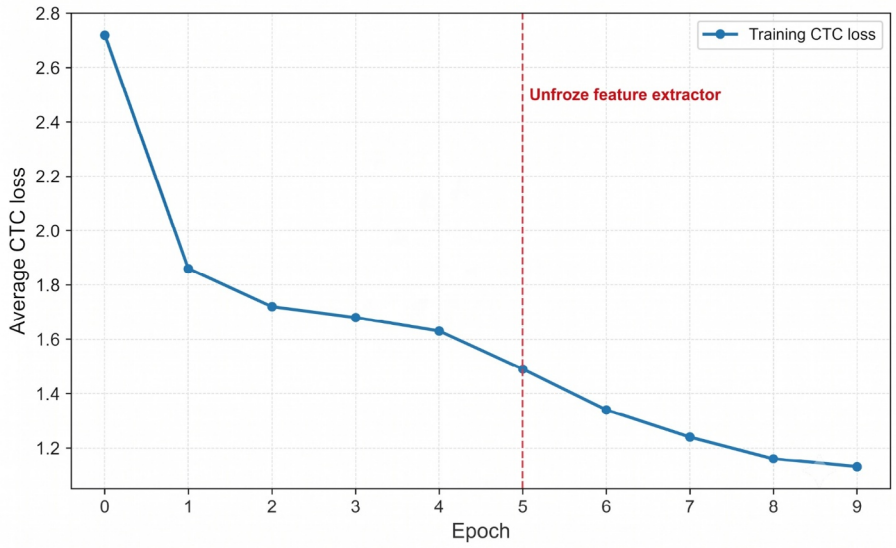


Figure 3. Sinhala ASR model: training loss convergence (epochs 0–9)
Abbreviation: CTC: Connectionist temporal classification.

6.1. Evaluation objectives

The evaluation was designed around measurable indicators that reflect learning efficiency and system performance. Rather than relying on qualitative observations alone, the analysis focuses on quantifiable metrics derived from system logs and speech model outputs.

The primary evaluation endpoints include:

- Trials-to-mastery/time-to-mastery: defined as the number of attempts required for a learner to reach and maintain a predefined mastery threshold for a phoneme or skill (Segal *et al*., 2018; van de Sande, 2013).
- Retention performance: assessed through delayed post-test activities or optional follow-up sessions designed to determine whether previously acquired skills remain stable over time.
- Pronunciation recognition accuracy: evaluated using phoneme-level and word-level error metrics produced by the speech recognition model (Morris *et al*., 2004; Yang *et al*., 2022).

These metrics collectively allow the system to evaluate both instructional effectiveness and pronunciation analysis reliability.

6.2. Comparative conditions

To understand the effect of adaptive sequencing, the evaluation framework considered several instructional policies. The following sequencing strategies were compared:

- Fixed curriculum ordering, where training activities follow a predefined sequence without adaptation.
- Threshold-based staircase progression, where difficulty increases or decreases according to recent performance.
- Bandit-based adaptive sequencing, where next-item selection is determined using algorithms such as Thompson sampling, UCB, or contextual LinUCB (Segal *et al*., 2018).
- Bandit sequencing with prerequisite constraints and monitoring filters, where candidate task selection is restricted by pedagogical rules and behavioral monitoring signals (Segal *et al*., 2018; Seo, 2017).

This comparative structure allows the analysis to provide a framework for assessing whether adaptive sequencing can improve task allocation and learning efficiency in future evaluations.

6.3. Speech model evaluation

The pronunciation evaluation model was assessed using in-sample testing on 993 speech samples drawn from the constructed dataset. Since all recordings originated from only five speakers and no independent speaker-level partition was available, the reported phoneme error rate should be interpreted as an initial feasibility result rather than a measure of generalization performance. Future work will employ speaker-independent leave-one-speaker-out cross-validation to provide a more rigorous evaluation. The evaluation focused on phoneme-level recognition accuracy and the effectiveness of the pronunciation scoring pipeline.

$$ER = \frac{S + D + I}{N} \tag{8}$$

Phoneme recognition performance was evaluated using the phoneme error rate, which measures the proportion of substitution (S), deletion (D), and insertion (I) errors relative to the total number of reference phonemes (N) (Morris *et al*., 2004), as shown in **Equation 8**. This metric was derived from the standard error rate formulation used in automatic speech recognition evaluation (Morris *et al*., 2004).

As shown in Table 4, the observed average phoneme error rate of 0.289 indicates moderate phoneme recognition performance under the conditions of the dataset. Considering that the recordings were collected from hearing-impaired speakers, this result suggests that the model can capture phonetic patterns despite the variability in articulation. Average word similarity does not represent classification accuracy. Rather, it measures normalized sequence similarity between predicted and reference phoneme sequences using Python's SequenceMatcher algorithm. The value is intentionally conservative because insertions, deletions, and substitutions all reduce the similarity score. Phoneme error rate and word similarity therefore evaluate different aspects of model behavior and are not directly comparable.

The evaluation primarily focused on phoneme-level recognition because the objective of the proposed system is pronunciation assessment rather than isolated-word recognition. In speech therapy, identifying incorrectly articulated phonemes is more clinically informative than

Table 4. Sample testing results

Metric	Value
Total samples evaluated	993
Average phoneme error rate	0.289
Average word similarity	0.229
Final CTC loss	1.1343

Abbreviation: CTC: Connectionist temporal classification.

 https://doi.org/10.6977/IJoSI.202608_10(4).0006

determining whether the entire word was recognized correctly. Phoneme-level evaluation enables the system to generate targeted goodness-of-pronunciation scores and individualized feedback for specific speech sounds, directly supporting therapeutic intervention. Word-level similarity is included as a preliminary validation step to reject severely mismatched utterances before detailed phoneme analysis, whereas phoneme error rate serves as the principal performance metric for evaluating pronunciation quality.

The final CTC loss value of 1.1343 further confirms stable convergence during model training. This result aligns with the behavior observed during the dual-phase fine-tuning process, where loss reduction accelerated after the feature encoder was unfrozen.

6.4. Word-level error analysis

To better understand model behavior across different phonetic structures, phoneme error rates were analyzed for each target word in the dataset. As illustrated in Figure 4, simpler stop-based words such as ibba (W2) exhibited lower error rates compared to words containing consonant clusters, such as kridawa (W20). This observation is consistent with established phonetic research, which shows that consonant clusters require more complex articulatory coordination than single consonant structures (Lee & Tsoi, 2021; Nonis, 2015).

For hearing-impaired speakers, these articulatory challenges can lead to phoneme substitutions, deletions, or distortions, particularly when multiple consonants occur consecutively within a word. This pattern suggests that cluster-heavy words may require additional training emphasis within the adaptive curriculum.

6.5. Adaptive behavior and task allocation

Beyond speech recognition accuracy, the platform's adaptive engine was evaluated through analysis of task allocation patterns and learner interaction logs.

Adaptive behavior was assessed by examining how training activities are distributed across learners and whether the system shifts task selection toward weaker phoneme categories over time (Segal *et al.*, 2018; van de Sande, 2013). Key indicators included:

- non-uniform distribution of task exposure across learners,
- increased practice frequency for low-mastery phonemes,
- gradual reduction of repeated exposure to already-mastered skills,
- policy auditing procedures monitor how often each candidate task is selected and whether exploration decreases as the learner model becomes more confident (Segal *et al.*, 2018).

6.6. Feedback system effectiveness

The pronunciation evaluation pipeline produces normalized confidence scores that are mapped to user feedback messages. The system categorizes pronunciation quality into several ranges:

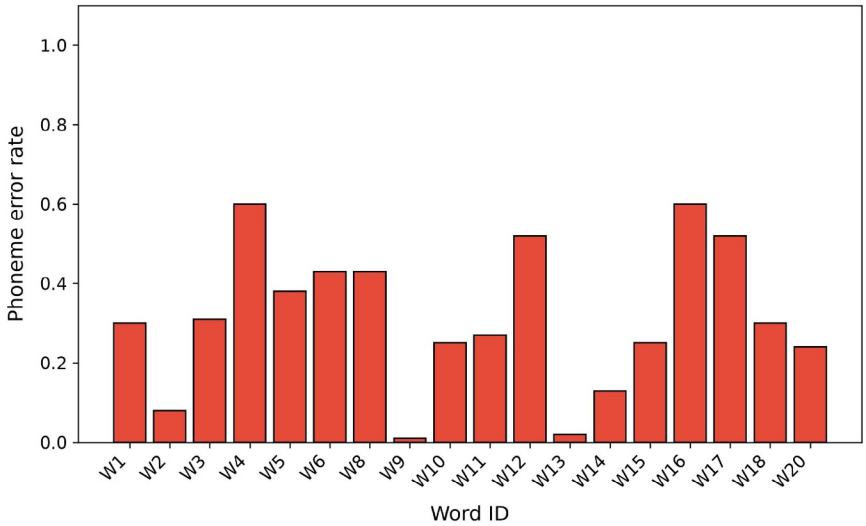


Figure 4. Phoneme error rate per word (W)

- High-confidence pronunciations (score ≥ 0.75) generate positive reinforcement messages.
- Moderate scores (0.35–0.55) trigger corrective guidance indicating that pronunciation should be improved.
- Very low similarity scores result in immediate retry prompts, preventing severely incorrect pronunciations from progressing to detailed evaluation (Andrijašević & Vukelić, 2024; Witt & Young, 2000; Yang *et al*., 2022).

This score-to-feedback mapping ensures that learners receive clear and actionable pronunciation guidance while maintaining a supportive training environment.

It should be noted that the dataset used in this study was relatively small, consisting of recordings from only five hearing-impaired participants collected during clinical therapy sessions. The limited availability of labeled Sinhala speech data from hearing-impaired children posed a challenge for large-scale training and evaluation. Therefore, the reported results should be interpreted as preliminary findings demonstrating the feasibility of the proposed pronunciation coaching framework.

6.7. Mastery-convergence and difficulty matching

The proposed learner model was designed such that mastery probabilities are expected to follow a characteristic convergence pattern. Early training stages produce rapid updates as the system gathers evidence about learner abilities. Over time, mastery estimates stabilize as more interaction data becomes available (van de Sande, 2013).

Adaptive policies such as staircase adjustment and bandit-based sequencing aim to maintain practice near an individualized difficulty range. This is reflected in reduced exposure to tasks that are either excessively easy or overly difficult (Segal *et al*., 2018; Seo, 2017).

The proposed adaptive architecture is intended to dynamically align tasks with learner performance, supporting more efficient auditory training.

6.8. Algorithm selection and model suitability

The learner modeling algorithms were selected to balance interpretability, data efficiency, and pedagogical safety within a low-resource clinical setting. As summarized in Table 3, the architecture distributes adaptive decision-making across modules rather than relying on a single complex model. Module M0 uses rubric-based pre-linguistic profiling because early learner interactions provide too little data for clustering or deep representation models to be reliable or clinically interpretable. Module M1 applies CAT-style difficulty adjustment since more advanced approaches such as item response theory and deep knowledge tracing require larger calibrated item banks and longer response histories than are available in the present Sinhala auditory training context. Module M2 adopts prerequisite-aware recommendation and weighted ranking instead of purely reinforcement-learning-based sequencing, as reward-driven policies may ignore pedagogical ordering and prerequisite dependencies. Module M3 uses rule-guided behavioral monitoring and scaffolded intervention because fully learned intervention policies may produce unsafe or poorly timed feedback in therapy-oriented settings. Module M4 performs weighted risk prediction rather than deep neural or ensemble modeling, since the available dataset is limited and more complex predictors would reduce interpretability while increasing overfitting risk. Overall, the selected methods were preferred not for algorithmic sophistication alone, but because they provide a more appropriate combination of transparency, robustness, and clinical controllability than more data-intensive black-box alternatives.

7. Conclusion

This study presented an integrated adaptive auditory training platform designed to support hearing-impaired Sinhala-speaking children. By combining an intelligent tutoring framework with an automated pronunciation evaluation (Fernando *et al*., 2024; Gnadlinger *et al*., 2024) module, the system demonstrates how modern speech technology and adaptive learning algorithms can be used to extend the accessibility of AVT.

The results indicate that integrating clinical speech data collection with a dual-phase Wav2Vec 2.0 fine-tuning strategy enables the development of a reliable pronunciation evaluation model for a low-resource language environment. The staged fine-tuning approach allowed the model to preserve the robustness of pretrained speech representations while adapting to the acoustic characteristics of hearing-impaired speech. The resulting system achieved stable convergence during training and produced phoneme-level recognition performance sufficient for pronunciation feedback generation.

Beyond pronunciation evaluation, the adaptive tutoring architecture enables the platform to personalize auditory training activities based on estimated learner mastery. By combining Bayesian knowledge tracing with bandit-based task selection policies, the system dynamically adjusts training sequences to maintain an appropriate difficulty level for each learner. This adaptive behavior supports personalized practice through adaptive task sequencing. The effectiveness of this approach will be evaluated in future user studies.

However, access to therapy often depends on therapist availability and frequent in-person sessions. The proposed system addresses this limitation by enabling structured self-learning sessions that can be performed outside clinical environments while still preserving therapy-oriented training principles.

Future work will involve expanding the dataset to include a larger and more diverse population of speakers and evaluating the system using out-of-sample testing scenarios. Additional improvements may include optimizing the pronunciation model for deployment as a lightweight mobile application, enabling home-based rehabilitation through mobile devices, and increasing the accessibility of auditory training for hearing-impaired children.

Acknowledgments

None.

Funding

None.

Conflict of interest

The authors declare they have no competing interests.

Author contributions

Conceptualization: Thilina Lakshan Samarasekara
Formal analysis: Himakara Liyanage
Investigation: Samanthi Eranga Siriwardhana, Lokesha Weerasinghe
Methodology: Thulasika Nayananjalee Weerasinghe, Thinama Renugi Wijesekara
Writing–original draft: Thilina Lakshan Samarasekara, Thinama Renugi Wijesekara
Writing–review & editing: Samanthi Eranga Siriwardhana, Lokesha Weerasinghe

Availability of data

The dataset used in this study is publicly available on Kaggle (https://www.kaggle.com/datasets/thilinal5/sinhala-pronunciation-evaluation-metrices) and can be obtained from the authors upon request.

References

Andrijašević, A., & Vukelić, B. (2024). Generating Speech Material for Auditory Training Exercises Using ChatGPT Chatbot. In: *Proceedings of the 2024 47th MIPRO ICT and Electronics Convention (MIPRO),* May 20-24, 2024, Opatija, Croatia. 126-131.

https://doi.org/10.1109/MIPRO60963.2024.10569423

Babu, A., Wang, C., Tjandra, A., Lakhotia, K., Xu, Q., Goyal, N., Singh, K., von Platen, P., Saraf, Y., Pino, J., Baevski, A., Conneau, A., & Auli, M. (2022). XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale. *arXiv.* arXiv:2111.09296.

https://doi.org/10.21437/interspeech.2022-143

Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations (Version 3). *Advances in neural information processing systems, 33,* 12449-12460.

https://doi.org/10.48550/arXiv.2006.11477

Colibaba, S., Gheorghiu, I., Colibaba, A., Ursa, O., Antoniță, C., & Cîrșmari, R. (2022, November 17–19). The Voice Project: Habilitating hearing-impaired children to recover hearing and lead a normal life. In: *Proceedings of the 2022 E-Health and Bioengineering Conference (EHB),* Online. 1-4.

https://doi.org/10.1109/EHB55594.2022.9991663

Corbett, A. T., & Anderson, J. R. (1994). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modelling and User-Adapted Interaction*, *4*(4), 253-278.

https://doi.org/10.1007/BF01099821

Fernando, S., Leo, C., Tharushika, C., Mawaththa, R., Perera, J., Vidhanaarachchi, S., & Kasthurirathna, D. (2024). Assisting Hearing Impaired Children by Personalized Gamification of Auditory Verbal Therapy. In: *Proceedings of the 2024 6th International Conference on Advancements in Computing (ICAC),* December 12-13, 2024, Colombo, Sri Lanka. 414-419.

https://doi.org/10.1109/ICAC64487.2024.10850970

Gnadlinger, F., Werminghaus, M., Selmanagić, A., Filla, T., Richter, J. G., Kriglstein, S., & Klenzner, T. (2024). Incorporating an intelligent tutoring system into a game-based auditory rehabilitation training for adult cochlear implant recipients: Algorithm development and validation. *JMIR Serious Games*, *12*(1), e55231.

https://doi.org/10.2196/55231

Graves, A., Fernández, S., Gomez, F., & Schmidhuber, J. (2006). Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In: *Proceedings of the 23rd International Conference on Machine Learning,* June 25-29, 2006, Pittsburgh, PA. 369-376

https://doi.org/10.1145/1143844.1143891

Ishihara, M., & Tsuda, M. (2021). Hearing support system for the hearing impaired. In: *Proceedings of the 2021 IEEE 3rd Global Conference on Life Sciences and Technologies (LifeTech),* March 9-11, 2021, Nara, Japan. 473-474.

https://doi.org/10.1109/LifeTech52111.2021.9391875

Lee, W. S., & Tsoi, I. C. Y. (2021, January 24-27). Acoustical Characteristics of the Cantonese Vowels and Tones

Produced by Hearing Impaired Speakers. In: *Proceedings of the 2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP),* Online. 1-5.

https://doi.org/10.1109/ISCSLP49672.2021.9362118

Liu, C., & Fu, Q. J. (2007). Estimation of vowel recognition with cochlear implant simulations. *IEEE Transactions on Biomedical Engineering*, *54*(1), 74-81.

https://doi.org/10.1109/TBME.2006.883800

Morris, A. C., Maier, V., & Green, P. D. (2004). From WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition. In: *Proceedings of InterSpeech 2004,* October 4-8, 2004, Jeju Island, Korea. 2765–2768.

https://doi.org/10.21437/Interspeech.2004-668

Nonis, P. D. M., & Hettiarachchi, S. (2014). A study of phonetic and phonological development of Sinhala speaking children in the Puttalam District aged 3; 0-3; 11 years. *Journal of the Faculty of Graduate Studies*, *3*, 8-25.

https://doi.org/10.13140/RG.2.1.1152.0486

Segal, A., Ben David, Y., Williams, J. J., Gal, K., & Shalom, Y. (2018). Combining difficulty ranking with multi-armed bandits to sequence educational content. In: *Proceedings of the International Conference on Artificial Intelligence in Education,* June 27-30, 2018, London, United Kingdom. 317-321).

https://doi.org/10.1007/978-3-319-93846-2_59

Seo, D. G. (2017). Overview and current management of computerized adaptive testing in licensing/certification examinations. *Journal of Educational Evaluation for Health Professions*, *14*, 17.

https://doi.org/10.3352/jeehp.2017.14.17

van De Sande, B. (2013). Properties Of The Bayesian Knowledge Tracing Model. *Journal of Educational Data Mining*, *5*(2), 1-10.

https://doi.org/10.5281/zenodo.3554629

Weerasinghe, R., Wasala, A., & Gamage, K. (2005). A rule based syllabification algorithm for Sinhala. In: *Proceedings of the International Conference on Natural Language Processing,* October 11-13, 2005, Jeju Island, Korea. 438-449.

https://doi.org/10.1007/11562214_39

Witt, S. M., & Young, S. J. (2000). Phone-level pronunciation scoring and assessment for interactive language learning. *Speech Communication*, *30*(2-3), 95-108.

https://doi.org/10.1016/S0167-6393(99)00044-8

Yang, M., Hirschi, K., Looney, S. D., Kang, O., & Hansen, J. H. L. (2022). Improving Mispronunciation Detection with Wav2vec2-based Momentum Pseudo-Labeling for Accentedness and Intelligibility Assessment. In: *Proceedings of Interspeech 2022,* September 18-22, 2022, Incheon, Korea. 4481–4485.

https://doi.org/10.21437/interspeech.2022-11039

Yong-Xin, L., Shuang, L., & De-Min, H. (2007). Research advances in post-operative rehabilitation following cochlear implant. *Journal of Otology*, *2*(2), 92-96.

https://doi.org/10.1016/S1672-2930(07)50019-X