

**ADAPTIVE ENGLISH LANGUAGE LEARNING  
PLATFORM FOR AUTISTIC CHILDREN IN  
SRI LANKA**

**INDIVIDUAL THESIS**

Navindu Gimhan Samaranayake

IT22266378

Phoneme-aware Adaptive Pronunciation Support Component

R26-IT-006

Report submitted in partial fulfilment of the requirements for the

B.Sc. (Hons) Degree in Information Technology

Department of Computing

Sri Lanka Institute of Information Technology

Sri Lanka

October 2026

## DECLARATION AND SUPERVISOR CERTIFICATION

**Candidate declaration.** I declare that this thesis is my own work and does not incorporate, without acknowledgement, material previously submitted for a degree or diploma at any university or institute of higher learning. To the best of my knowledge, it contains no material previously published or written by another person except where acknowledgement is made in the text. I grant the Sri Lanka Institute of Information Technology the non-exclusive right to reproduce and distribute this thesis, in whole or in part, in print, electronic, or other media, while retaining the right to use the content in future work.

**Academic integrity and AI disclosure.** I have disclosed all use of generative AI and other automated tools in accordance with the module policy. I remain responsible for the accuracy, originality, citations, code, data, analyses, and conclusions in this thesis.

Student name: Navindu Gimhan Samaranayake

Student ID: IT22266378

Signature: \_\_\_\_\_ Date: 11/10/2026

**Supervisor certification.** The above candidate carried out this research for the bachelor's degree thesis under my supervision.

| Role and name                    | Signature | Date |
|----------------------------------|-----------|------|
| Supervisor: Ms. Manori Gamage    |           |      |
| Co-Supervisor: Ms. Suriya Kumari |           |      |

## ABSTRACT

Children with Autism Spectrum Disorder (ASD) in Sri Lanka struggle with English pronunciation because English is taught as a second language through methods designed for typically developing students. Computer-Assisted Pronunciation Training (CAPT) systems assume neurotypical or adult learners, and no existing system integrates autism-sensitive practice, automated scoring, adaptive progression, and teacher monitoring. Using a design and development research approach, this study built an adaptive English pronunciation support component for Grade 3 autistic learners. A React Native application delivers picture prompts, reference audio, and phoneme-level guidance, followed by reinforcement activities during which speech is recorded. As proposed, recordings are compared with UK reference pronunciations using Mel-frequency cepstral coefficients (MFCC) and dynamic time warping (DTW). Because this proved unreliable for child voices, a wav2vec2 Goodness of Pronunciation (GOP) model and Whisper word verification were added. To adapt scoring to autistic speech, classroom recordings from a Sri Lankan special-education school were processed by a pipeline that separates child utterances from teacher prompts using pitch clustering and prompt-repeat alignment, yielding 167 child utterances. After blind teacher labelling, a small, supervised model trained on acoustic and phoneme features learns to predict teacher scores, evaluated with student-held-out cross-validation. An adaptive engine selects the next word from the weakest phoneme, word difficulty, and recurring weak sounds, and a teacher's dashboard provides scores, playback, and review of low-confidence results. Across 249 in-app attempts from 16 students, system scores correlated with teacher scores at  $r = 0.73$  (mean absolute error 7.9 points), while MFCC-DTW alone reached only  $r = 0.19$ . The individual contribution is the design, implementation, and evaluation of this component, including its scoring cascade, child-speech adaptation model, adaptive word selection, and teacher-review loop.

**Keywords:** pronunciation assessment, autism spectrum disorder, Goodness of Pronunciation, child speech adaptation

## **ACKNOWLEDGEMENTS**

I would like to express my sincere gratitude to my research supervisor, Ms. Manori Gamage, for her continuous guidance, valuable feedback, and support throughout the development of this research proposal. Her insights and academic expertise have been instrumental in shaping the direction of this research. I would also like to extend my sincere appreciation to my co-supervisor, Mrs. Suriya Kumari, for her guidance, encouragement, and constructive suggestions that contributed significantly to improving this research work.

I would also like to thank the Sri Lanka Institute of Information Technology (SLIIT) for providing the academic environment and resources necessary for conducting this research. Special appreciation is extended to my research group members for their collaboration and valuable discussions that helped refine the overall system concept.

Finally, I would like to express my heartfelt gratitude to my family and friends for their encouragement, patience and continuous support throughout my academic journey.

## TABLE OF CONTENTS

|                                                     |    |
|-----------------------------------------------------|----|
| 1 INTRODUCTION .....                                | 8  |
| 1.1 Overall project context.....                    | 9  |
| 1.2 Individual component context.....               | 9  |
| 1.3 Problem statement.....                          | 10 |
| 1.4 Research gap .....                              | 11 |
| 1.5 Aim and objectives.....                         | 12 |
| 1.6 Research questions or hypotheses .....          | 13 |
| 1.7 Scope limitations and assumptions .....         | 14 |
| 1.8 Individual contribution.....                    | 15 |
| 1.9 Thesis structure .....                          | 17 |
| 2 LITERATURE REVIEW AND RELATED WORK .....          | 17 |
| 2.1 Review method and scope.....                    | 18 |
| 2.2 Domain and theoretical foundations .....        | 18 |
| 2.3 Related research .....                          | 21 |
| 2.4 Existing systems products and techniques .....  | 22 |
| 2.5 Gap synthesis and conceptual framework.....     | 24 |
| 3 RESEARCH METHODOLOGY.....                         | 25 |
| 3.1 Research design.....                            | 26 |
| 3.2 Data participants or study materials.....       | 26 |
| 3.3 Experimental or study procedure .....           | 28 |
| 3.4 Measures and evaluation metrics .....           | 32 |
| 3.5 Data analysis .....                             | 33 |
| 3.6 Ethics privacy security and responsible AI..... | 34 |
| 3.7 Reproducibility and research artefacts .....    | 36 |

|                                                                       |    |
|-----------------------------------------------------------------------|----|
| 4 PRODUCT AND SOLUTION DESIGN.....                                    | 37 |
| 4.1 Requirements .....                                                | 38 |
| 4.2 Component architecture within the complete solution .....         | 40 |
| 4.3 Design alternatives and decisions .....                           | 42 |
| 4.4 Detailed design.....                                              | 43 |
| 4.5 Standards and professional practice .....                         | 50 |
| 5 IMPLEMENTATION INTEGRATION AND TESTING .....                        | 52 |
| 5.1 Development environment and technology stack .....                | 53 |
| 5.2 Component implementation .....                                    | 54 |
| 5.3 Integration with the other components.....                        | 54 |
| 5.4 Verification and software testing .....                           | 55 |
| 5.5 Deployment and operation .....                                    | 57 |
| 6 RESULTS AND VALIDATION.....                                         | 58 |
| 6.1 Research results.....                                             | 59 |
| 6.2 Product and technical results.....                                | 62 |
| 6.3 User or stakeholder validation .....                              | 64 |
| 6.4 Integration results.....                                          | 64 |
| 6.5 Summary of findings.....                                          | 65 |
| 7 DISCUSSION .....                                                    | 65 |
| 7.1 Interpretation of findings .....                                  | 66 |
| 7.2 Comparison with literature baselines and existing solutions ..... | 67 |
| 7.3 Research contribution .....                                       | 67 |
| 7.4 Product contribution.....                                         | 68 |
| 7.5 Limitations and threats to validity .....                         | 69 |
| 7.6 Implications.....                                                 | 70 |
| 8 CONCLUSION AND FUTURE DIRECTIONS .....                              | 71 |

|                                                           |    |
|-----------------------------------------------------------|----|
| 8.1 Conclusion .....                                      | 72 |
| 8.2 Achievement of objectives .....                       | 73 |
| 8.3 Individual research and product contributions .....   | 73 |
| 8.4 Future directions .....                               | 74 |
| REFERENCES.....                                           | 76 |
| GLOSSARY.....                                             | 77 |
| APPENDICES .....                                          | 78 |
| Appendix A Research instruments or extended results ..... | 79 |
| Appendix B Academic integrity and AI use disclosure ..... | 79 |

## LIST OF FIGURES

|                                                                                                    |           |
|----------------------------------------------------------------------------------------------------|-----------|
| <i>Figure 2.1: Conceptual framework of the Adaptive English Pronunciation Support Module .....</i> | <i>16</i> |
| <i>Figure 4.2: The component within the complete solution.....</i>                                 | <i>32</i> |
| <i>Figure 4.4: Session flow.....</i>                                                               | <i>35</i> |
| <i>Figure 4.5: Scoring cascade .....</i>                                                           | <i>37</i> |
| <i>Figure 4.6: Next-word decision .....</i>                                                        | <i>38</i> |
| <i>Figure 4.7: Entity–relationship view .....</i>                                                  | <i>39</i> |
| <i>Figure 4.8: Adaptation pipeline.....</i>                                                        | <i>41</i> |

## LIST OF TABLES

|                                                                                                                                                           |    |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------|----|
| <i>Table 2.1: Existing applications compared with the proposed component .....</i>                                                                        | 14 |
| <i>Table 2.2: Pronunciation scoring techniques .....</i>                                                                                                  | 15 |
| <i>Table 2.3: How existing work covers the requirements of this component.....</i>                                                                        | 15 |
| <i>Table 3.2: Evaluation metrics .....</i>                                                                                                                | 24 |
| <i>Table 3.3: Research artefacts .....</i>                                                                                                                | 28 |
| <i>Table 4.1: Functional requirements and how they were met.....</i>                                                                                      | 30 |
| <i>Table 4.2: Non-functional requirements .....</i>                                                                                                       | 31 |
| <i>Table 4.3: Design alternatives and decisions .....</i>                                                                                                 | 34 |
| <i>Table 5.1: Technology stack of the Pronunciation Support component .....</i>                                                                           | 44 |
| <i>Table 5.2: Automated tests for the component.....</i>                                                                                                  | 46 |
| <i>Table 5.3: Defects found and fixed during verification .....</i>                                                                                       | 47 |
| <i>Table 6.1: Attempts and teacher reviews by population .....</i>                                                                                        | 50 |
| <i>Table 6.2: Agreement between the system's overall score and teacher scores (n = 32)</i><br><i>.....</i>                                                | 50 |
| <i>Table 6.3: Layer-1 outcomes.....</i>                                                                                                                   | 51 |
| <i>Table 6.4: Layer-1 agreement with teacher scores .....</i>                                                                                             | 51 |
| <i>Table 6.5: Sweep of the Layer-1 accept threshold (teacher-reviewed attempts, where</i><br><i>"correct" means a teacher score of at least 70) .....</i> | 51 |
| <i>Table 6.7: Adaptation model, held-out evaluation (grouped by session) .....</i>                                                                        | 53 |
| <i>Table 6.8: Status of each objective.....</i>                                                                                                           | 54 |
| <i>Table 6.9: Technical results .....</i>                                                                                                                 | 54 |
| <i>Table 8.1: Achievement of objectives .....</i>                                                                                                         | 64 |

## LIST OF ABBREVIATIONS

| Abbreviation | Meaning                                           |
|--------------|---------------------------------------------------|
| AAC          | Augmentative and Alternative Communication        |
| AI           | Artificial Intelligence                           |
| API          | Application Programming Interface                 |
| ASD          | Autism Spectrum Disorder                          |
| ASR          | Automatic Speech Recognition                      |
| CAPT         | Computer-Assisted Pronunciation Training          |
| CI           | Confidence Interval                               |
| CORS         | Cross-Origin Resource Sharing                     |
| DNN          | Deep Neural Network                               |
| DTW          | Dynamic Time Warping                              |
| ESL          | English as a Second Language                      |
| F0           | Fundamental Frequency                             |
| FR           | Functional Requirement                            |
| GB           | Gigabyte                                          |
| GOP          | Goodness of Pronunciation                         |
| GPU          | Graphics Processing Unit                          |
| HMM          | Hidden Markov Model                               |
| HTTP         | Hypertext Transfer Protocol                       |
| IEC          | International Electrotechnical Commission         |
| IEEE         | Institute of Electrical and Electronics Engineers |
| ISO          | International Organization for Standardization    |

| Abbreviation | Meaning                                     |
|--------------|---------------------------------------------|
| JSON         | JavaScript Object Notation                  |
| JSONB        | Binary JSON (PostgreSQL data type)          |
| JWT          | JSON Web Token                              |
| L2           | Second Language                             |
| MAE          | Mean Absolute Error                         |
| MB           | Megabyte                                    |
| MFCC         | Mel-Frequency Cepstral Coefficient          |
| NFR          | Non-Functional Requirement                  |
| ORM          | Object-Relational Mapping                   |
| OWASP        | Open Worldwide Application Security Project |
| REST         | Representational State Transfer             |
| RFC          | Request for Comments                        |
| RQ           | Research Question                           |
| SDK          | Software Development Kit                    |
| SNR          | Signal-to-Noise Ratio                       |
| SVG          | Scalable Vector Graphics                    |
| UK           | United Kingdom                              |
| WAV          | Waveform Audio File Format                  |
| WCAG         | Web Content Accessibility Guidelines        |

# 1 INTRODUCTION

## 1.1 Overall project context

Children with Autism Spectrum Disorder (ASD) often find language learning difficult because of differences in communication, attention, sensory sensitivity and learning pace [1], [2]. Autism is not rare in Sri Lanka: a community screening study found a prevalence of about 1 in 93 among children aged 18–24 months [3]. In Sri Lanka the difficulty is greater because English is taught as a second language (ESL) [4], and the classrooms, materials and teaching methods used for it are largely designed for typically developing students. Special education teachers also tend to manage several children with very different abilities, so it is hard to give each child structured, repeated, and individually paced practice.

The project Adaptive English Language Learning Platform for Autistic Children in Sri Lanka responds to this need with a teacher-supported tablet platform that adapts English learning activities to each autistic learner. The platform has several modules. This thesis covers only one of them: the Adaptive English Pronunciation Support Module.

## 1.2 Individual component context

The Adaptive English Pronunciation Support Module helps Grade 3 autistic children learn foundational English pronunciation and helps their teachers monitor and guide that learning. It is a React Native (Expo) application connected to a Node.js/Express backend with a PostgreSQL database.

A session starts with the teacher. The teacher chooses alphabet mode or word mode, chooses a vocabulary category aligned with the Grade 3 syllabus (animals, fruits, classroom objects or everyday actions), and chooses the first target word. The child then gets multimodal support before any assessment takes place: a picture prompt, the reference pronunciation as audio, and phoneme-level guidance. These supports can be repeated as often as needed. Reinforcement activities follow: listen-and-choose, a

sound-segmentation activity that needs no speech, and spoken-word practice in which the child's pronunciation is recorded.

Each recording is scored on the server. As proposed, the child's speech is compared with a UK English reference pronunciation using Mel-frequency cepstral coefficients (MFCC) [5] and dynamic time warping (DTW) [6]. During development this word-level comparison proved unreliable for children's voices, so it was extended with phoneme-level Goodness of Pronunciation (GOP) scoring [7] using a wav2vec2 model [8], which also identifies the weakest sound in the word, and with Whisper-based [9] checking that the target word was actually spoken. The child sees only simple, encouraging feedback, never a numerical score. The teacher sees detailed results: word-level scores, the weakest phoneme, recording playback, session history, and a review queue that flags low-confidence results for teacher judgement. After each attempt, an adaptive mechanism selects the next word using the child's score, the weakest phoneme, word difficulty and recurring difficulty patterns.

To adapt the scoring to the target population, the component also includes a child-speech adaptation model. Classroom recordings of autistic students, made in a teacher-prompt and child-repetition format, are processed automatically. Speech segments are found, assigned to the teacher or the child by voice pitch, and the child's utterances are cut into individual clips. After blind labelling by a teacher, who cannot see the system's scores, a small supervised model learns to predict the teacher's score from the measurements the scoring pipeline already produces: phoneme-level GOP scores, DTW distance, word verification and audio quality. The model is adopted only if it predicts teacher judgement more accurately than the existing pipeline on children it has not seen.

### **1.3 Problem statement**

Autistic children in Sri Lankan special education settings need structured, repetitive and individualized pronunciation practice, but the tools and methods available for English pronunciation are designed for typically developing learners and adults. Teachers judge pronunciation mostly by ear, so assessment is subjective, varies

between sessions and teachers, and is hard to record over time [10]. Because special education teachers manage several children with different abilities, they cannot reliably track which sounds each child repeatedly finds difficult or choose practice words that target those sounds.

Existing digital pronunciation tools cannot fill this gap. Their speech assessment models are built and tuned on adult or typically developing speakers [10], [11]. They use drill-based interaction that does not suit autistic learners. They often show detailed scores directly to the learner, which can discourage children who are sensitive to failure [12]. They also fail to give teachers the evidence they need to make instructional decisions. There is therefore a need for a pronunciation support component that is sensitive to autistic learners, scores children's pronunciation objectively and consistently, adapts practice to each child's phoneme-level difficulties, and keeps the teacher in control of interpreting results.

#### **1.4 Research gap**

Research on technology for autistic learners and research on Computer-Assisted Pronunciation Training (CAPT) have developed largely separately. Technology-based interventions for children with ASD have shown promise [13], [14], but they focus mainly on vocabulary, communication and social skills, and pronunciation receives little attention. CAPT and automatic pronunciation assessment, by contrast, are technically mature [12], [10], but almost all of that work assumes neurotypical or adult learners. Working on children's pronunciation assessment shows that child speech remains harder for automatic systems than adult speech [15], [11].

The building blocks exist separately: visual scaffolding suits autistic learners [16], MFCC and DTW can measure pronunciation similarity [5], [6], phoneme-level GOP scoring can locate individual mispronounced sounds [7] and teacher dashboards can support teachers' decisions. They have not been combined into one system in which:

1. pronunciation practice is embedded in autism-sensitive reinforcement activities rather than isolated recording drills

2. scoring works at the level of individual phonemes, so it shows which sound is difficult rather than only whether the word was correct
3. the next practice word is chosen from the child's recurring phoneme-level difficulties rather than from a single pass/fail score
4. Teachers receive detailed evidence, review low-confidence results, and remain the final judges of a child's performance.

The gap is especially wide in Sri Lanka. Grade 3 is when formal ESL instruction begins [4], assistive technology in autism classrooms is limited, and local speech technology research has concentrated on Sinhala rather than English pronunciation support.

## **1.5 Aim and objectives**

### **Aim**

To design, develop and evaluate an adaptive pronunciation support component that helps Grade 3 autistic children learn foundational English pronunciation through multimodal practice, automated pronunciation assessment and personalized word progression, while enabling teachers to monitor and guide each child's progress.

### **Objectives**

1. To collect and organize Grade 3 English learning content suitable for autistic learners: the alphabet, categorized vocabulary, UK English reference pronunciations, images and activity materials.
2. To design and develop a multimodal learning interface that supports pronunciation through picture prompts and audio playback, following autism-friendly design principles.
3. To implement speech acquisition and preprocessing that captures pronunciation attempts on a tablet device, checks recording quality and prepares the audio for analysis.
4. To develop a pronunciation assessment mechanism that compares the child's speech with the expected pronunciation, at word level using MFCC and DTW

and at phoneme level using Goodness of Pronunciation scoring, and that produces word-level and phoneme-level scores.

5. To design reinforcement activities that let the child practise and show understanding of learned words, including activities that need no speech.
6. To implement an adaptive learning mechanism that selects the next word from the child's pronunciation performance, recurring phoneme-level difficulties, and word complexity.
7. To develop teacher-facing monitoring that provides word-level scores, weakest-phoneme information, recording playback, progress history, and review of low-confidence results.
8. To evaluate the component's functionality and the agreement between its pronunciation scores and teacher judgement.
9. To develop a child-speech adaptation model, trained on teacher-labelled recordings of autistic students, that adapts pronunciation scoring to the speech of the target population, and to evaluate it against the existing scoring pipeline on held-out students.

## **1.6 Research questions or hypotheses?**

- **RQ1.** Can word-level MFCC-DTW similarity against an adult UK English reference reliably assess the pronunciation of autistic children?
- **RQ2.** Does phoneme-level Goodness of Pronunciation scoring produce pronunciation scores that agree with teacher judgement for this population?
- **RQ3.** Can a pronunciation component that adapts to phoneme-level difficulty, and keeps teacher review in the loop, be implemented to provide consistent and supportive pronunciation practice for Grade 3 autistic learners?
- **RQ4.** Can a small supervised model trained on teacher-labelled recordings of autistic students improve agreement between automated pronunciation scores and teacher judgement compared with the unadapted scoring pipeline?

## 1.7 Scope limitations and assumptions

### Scope

- The component targets foundational English pronunciation for Grade 3 autistic learners in Sri Lankan special education settings.
- Content covers 78 target items with UK English reference recordings: the 26 letters of the alphabet and 52 words across four categories (animals, fruits, classroom objects and everyday actions).
- The component covers multimodal learning support, reinforcement activities, speech recording, automated pronunciation scoring, adaptive next-word selection and teacher-facing monitoring.

The interface is in English and Sinhala.

The component is an educational support tool. In line with the ethical commitments of the proposal, its outputs are learning indicators for teachers, not diagnostic or clinical assessments.

### Limitations

- **Adult reference voices.** Every reference pronunciation is an adult UK English recording. Children's voices differ acoustically from adult voices [17], which limits how well MFCC-DTW template matching works.
- The child-speech adaptation model is trained on a small set of classroom recordings from a single school. Its training labels come from teacher judgement, and teacher and child speech are separated automatically by pitch, so some clips may be misattributed and must be excluded during labelling. Its results therefore indicate feasibility rather than generalizable accuracy.
- **Not trained on children.** The phoneme recognition model is pretrained on general speech and has not been fine-tuned on child or autistic speech.

- **Reference recording quality varies.** An audit found that some reference recordings do not decode cleanly to their intended phonemes and need re-recording.
- **Small, non-random evaluation sample.** The evaluation involved a small number of students, and the teacher-reviewed attempts used to measure agreement were selected by uncertainty-based review [18], not at random. Results are therefore preliminary.
- **Recalibration not yet active.** The recalibration of scores against teacher reviews is implemented but runs in evidence-only mode until enough reviewed attempts are available for each learner population.
- **Network dependency.** Scoring runs on the server, so the application needs a network connection to the backend.

### **Assumptions**

- A teacher supervises every session and selects the starting category and word.
- Recordings are made in a reasonably quiet classroom on a device with a working microphone.
- The judgement of a trained special education teacher is a valid reference for the child's actual pronunciation.
- In the classroom recordings, the higher-pitched speaker responding after each teacher prompt is the child.
- Guardian consent is obtained before a child takes part, and child assent is considered, as set out in the ethical considerations.

### **1.8 Individual contribution**

The author designed, implemented and evaluated the Adaptive English Pronunciation Support Module end to end:

- **Content:** prepared the pronunciation content: target words and letters, categories, phoneme mappings, word difficulty profiles and UK English reference audio.
- **Tablet interface:** designed and built the session flow (mode, category and word selection, picture and audio support, listen-and-choose, the sound-segmentation activity, spoken-word recording) and child-friendly feedback that never shows numerical scores to the child.
- **Recording:** implemented on-device speech capture, plus a server-side recording-quality check that rejects recordings without enough usable speech.
- **Three-layer scoring:**
  - MFCC-DTW comparison against reference audio, restricted so it can only accept an attempt, never fail a child on its own.
  - wav2vec2 Goodness of Pronunciation scoring that identifies the weakest phoneme [7], [8].
  - Per-population recalibration against teacher-reviewed scores, validated by cross-validation and run in evidence-only mode until enough reviews exist for each population.
  - A child-speech adaptation model: a small, supervised model (ridge regression and isotonic recalibration [19]) that learns to predict the teacher's score from the pipeline's measurements. It is trained on classroom recordings of autistic students, which are separated into child utterances by voice pitch and labelled blind by teachers, and it is adopted only if it beats the existing scoring on held-out students.
- **Word verification:** used Whisper [9] to check that the target word was spoken, so low-confidence results are flagged for teacher review rather than presented as certain.

- **Adaptive next-word selection:** chooses easier or related practice words from the child's score, weakest phoneme, word difficulty and recurring difficulty patterns.
- **Teacher monitoring:** built word-level and phoneme-level results, recording playback, session history and an uncertainty-ranked review queue.
- **Child-speech adaptation:** built a pipeline that turns classroom recordings into labelled child pronunciation data (speech segmentation, pitch-based teacher/child separation, target-word alignment from the teacher's prompt, and a blind labelling tool). Also built a small supervised scoring model, ridge regression and isotonic recalibration evaluated by student-held-out cross-validation, that is adopted only if it improves the existing pipeline.
- **Evaluation:** compared system scores with teacher-reviewed scores.

## 1.9 Thesis structure

**Chapter 2** reviews the literature that shapes the component: autism and second-language learning, computer-assisted pronunciation training, speech representations such as MFCC and DTW, Goodness of Pronunciation scoring, self-supervised speech models, the acoustics of children's speech and methods for calibrating automated scores against human judgement. It ends by comparing existing work with the component's requirements and presenting the conceptual framework.

**Chapter 3** describes the research methodology: the design and development approach, the learning content and participants, the study procedure, the evaluation metrics, the data analysis, and the ethical, privacy and reproducibility measures.

**Chapter 4** presents the product and solution design, covering the requirements, the component's place in the complete platform, the main design alternatives and decisions, the detailed design of the session flow, scoring cascade, adaptive word selection, data and API, and the standards followed.

**Chapter 5** explains how the component was implemented, integrated with the other platform components, tested and deployed.

**Chapter 6** reports the results: agreement between the system's scores and teacher judgement, the technical results, stakeholder validation and integration results, followed by a summary of the findings against each research question.

**Chapter 7** discusses what the findings mean, compares them with the literature and existing solutions, sets out the research and product contributions, and examines the limitations and threats to validity.

**Chapter 8** concludes the thesis, reviews how far each objective was achieved, summarizes the individual contributions and suggests directions for future work. The references, glossary and appendices follow.

## 2 LITERATURE REVIEW AND RELATED WORK

### 2.1 Review method and scope

This narrative review organized around four questions the Pronunciation Support component has to answer:

1. What do autistic children need from pronunciation practice?
2. How can a machine judge pronunciation?
3. Why is children's speech harder to judge?
4. How can an automated judgement be made trustworthy enough for a teacher to act on?

Literature was drawn from IEEE Xplore, the ACL Anthology, the ISCA Archive (Interspeech), the ACM Digital Library, PubMed and Google Scholar, using combinations of these terms: *autism spectrum disorder*, *technology-based intervention*, *computer-assisted pronunciation training*, *automatic pronunciation assessment*, *goodness of pronunciation*, *mispronunciation detection*, *children's speech recognition*, *wav2vec 2.0*, *dynamic time warping* and *score calibration*.

#### **Included:**

- studies of technology-supported learning for autistic children or children with developmental disabilities
- methods for automatic pronunciation assessment at the word or phoneme level
- research on the acoustics and automatic processing of children's speech
- methods for calibrating automated scores with human judgement.

**Excluded:** work on sentence-level fluency, prosody-only assessment and reading comprehension, except where it informs word-level scoring. Existing applications were reviewed from their published descriptions and feature sets.

## 2.2 Domain and theoretical foundations

Autism Spectrum Disorder is a neurodevelopmental condition that affects communication and social interaction and is often accompanied by sensory sensitivities and a preference for routine [1]. Speech and prosody in autistic speakers can differ from those of typically developing peers [2]. For autistic children learning English as a second language, as in Sri Lankan schools where formal English instruction begins in Grade 3 [4], pronunciation underpins being understood, listening comprehension and taking part in class.

Two principles from special education shape a pronunciation tool for this population.

- **Structured, visual support.** Predictable routines and visual cues help autistic learners link sounds with meaning while reducing cognitive load [16]. A meta-analysis of technology-based interventions for ASD reported positive effects [13], mobile devices have been used successfully to teach children with developmental disabilities [14], and systematic reviews report benefits of mobile technology for autistic learners.
- **Errorless learning.** Learners are guided towards correct responses rather than corrected after failure, which reduces frustration and avoidance [20]. For pronunciation, this means the child receives encouragement while detailed diagnostic information goes to the teacher.

### Computer-Assisted Pronunciation Training

Computer-Assisted Pronunciation Training (CAPT) lets learners hear model speech, practise, and receive automated feedback. Rogerson-Revell [12] observes that CAPT technology has advanced faster than its pedagogy: much automated feedback is technically precise but not motivating for the learner. How feedback is delivered matters especially for children with special educational needs.

### Speech signal representation

- **Mel-frequency cepstral coefficients** (MFCCs) summarize the short-term spectral envelope on a perceptual frequency scale [5].
- **Dynamic time warping (DTW)** aligns two feature sequences spoken at different rates by finding the minimum-cost path between them [6].

Comparing a learner's MFCCs with a reference recording via DTW gives a word-level similarity score without any training data. DTW measures acoustic closeness to a particular reference speaker, so it is sensitive to differences in voice as well as in pronunciation.

### **Goodness of Pronunciation**

GOP compares the likelihood of the intended phoneme with that of the best competing phoneme over the same audio segment, normalized by the segment's duration [7], [21]. Rather than measuring closeness to a reference voice, GOP asks whether each sound is recognizable as the intended phoneme, so it can identify which phoneme in a word was mispronounced. Deep neural network acoustic models later improved GOP-based mispronunciation detection [22].

### **Self-supervised speech models and speech recognition**

wav2vec 2.0 learns speech representations from large amounts of unlabelled audio and can be fine-tuned with little labelled data [8]. Fine-tuned on phoneme transcriptions across many languages, it recognizes phonemes without a language-specific acoustic model, which makes it possible to compute GOP-style phoneme scores for any word with a phoneme transcription [23]. Whisper, trained on about 680,000 hours of weakly supervised audio, provides robust word recognition that can confirm whether the intended word was spoken [9].

### **Calibration and human-in-the-loop assessment**

An automated score is useful to a teacher when its value matches what the teacher would give.

- **Isotonic regression** learns a monotone mapping from model scores to human ratings and is a standard calibration method [19].

- **Ridge regression** combines several measurements into a stable prediction from small datasets [19], [24].
- **Uncertainty sampling** sends the most uncertain predictions to human reviewers first, making the best use of limited reviewing time [18], [25].

## 2.3 Related research

### 2.3.1 Automatic pronunciation assessment

El Kheir et al. [10] review automatic pronunciation assessment. They distinguish segmental assessment (phonemes, words) from supra-segmental assessment (stress, rhythm, intonation) and survey methods from GOP-based scoring to end-to-end neural models. Expert-annotated corpora such as speechocean762 [26] (5,000 English utterances from 250 non-native speakers, half of them children, scored at phoneme, word and sentence level) have enabled data-driven approaches. Recent models perform scoring and mispronunciation diagnosis together through multi-task learning, or use hierarchical transformers with pre-training, and interpretable models of L2 fluency have also been proposed. These models perform strongly but depend on large expert-annotated corpora of their target speakers.

### 2.3.2 Children's speech

Children's speech has higher fundamental and formant frequencies, longer and more variable segment durations, and greater variability than adult speech [17]. Speech recognition trained on adults therefore performs worse on children [11]. The NOCASA challenge was established because automatic assessment of non-native children's speech remains difficult and children's data is scarce [15]. For template matching such as MFCC-DTW, this means a child's correct pronunciation can still sit far from an adult reference template, which makes phoneme-identity measures such as GOP the more suitable basis for children.

### 2.3.3 Technology-supported language learning for autistic children

Technology research for autistic learners has focused mainly on vocabulary, communication and social skills. Augmentative and alternative communication (AAC) tools support non-speaking users but do not train pronunciation [27]. Pronunciation-focused interventions for autistic learners are few and narrow in scope, and their authors call for integrated automated speech assessment and adaptive feedback, the direction this component takes.

### 2.3.4 Adaptive sequencing

Most CAPT systems adapt on one aggregate score: the learner moves on after a pass and repeats after a fail. Phoneme-level scoring makes a richer signal available. When the same weakest phoneme recurs across several words, practice can target that sound in easier or related words, giving more individual and responsive progression.

## 2.4 Existing systems products and techniques

*Table 2.1: Existing applications compared with the proposed component*

| Application            | Main focus                              | Strengths                                            | Limitations for this context                                                                                                                                     |
|------------------------|-----------------------------------------|------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Google Read Along      | Speech-based reading tutor for children | Real-time feedback while reading aloud story library | Reading fluency rather than word pronunciation, no teacher-chosen words, phoneme diagnosis or adaptation to difficult sounds, not designed for autistic learners |
| ELSA Speak             | AI pronunciation coaching               | Phoneme-level feedback on English                    | Aimed at adult and adolescent L2 learners, scores shown directly to the learner, no autism-sensitive scaffolding or teacher monitoring                           |
| Duolingo               | Gamified language learning              | Engaging short exercises, some speaking tasks        | General learners, coarse speech feedback, no teacher oversight                                                                                                   |
| Proloquo2Go / Proloquo | AAC communication                       | Customizable symbol-based communication              | Communication support, not pronunciation training                                                                                                                |
| Avaz AAC               | AAC communication and learning          | Picture- and text-based AAC with progress analytics  | No comparison of the child's speech with a reference, no structured pronunciation                                                                                |

| Application        | Main focus                                                   | Strengths                                                                                                                                                         | Limitations for this context               |
|--------------------|--------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------|
|                    |                                                              |                                                                                                                                                                   | activities, no adaptation driven by speech |
| Proposed component | Adaptive pronunciation support for Grade 3 autistic learners | Multimodal practice, errorless activities, phoneme-level scoring, adaptation to the weakest phoneme, teacher review and monitoring, learns from teacher judgement | Research prototype evaluated in one school |

Table 2.2: *Pronunciation scoring techniques*

| Technique                                 | How it works                                               | Strengths                                                 | Considerations for autistic children's speech                                |
|-------------------------------------------|------------------------------------------------------------|-----------------------------------------------------------|------------------------------------------------------------------------------|
| MFCC + DTW                                | Aligns learner features with a reference recording         | No training data, fast, word-level                        | Sensitive to the child–adult voice difference, no phoneme diagnosis          |
| HMM/DNN GOP                               | Phoneme likelihood ratio over aligned segments             | Phoneme-level, interpretable                              | Needs an acoustic model for the language and speaker type                    |
| Self-supervised phoneme recognition + GOP | Phoneme posteriors from pretrained wav2vec 2.0             | No language-specific training, phoneme-level for any word | Pretrained mainly on adult speech                                            |
| End-to-end assessment models              | Neural networks trained on expert-scored utterances        | Strong accuracy on benchmark corpora                      | Need large annotated corpora of the target population                        |
| ASR word verification                     | Recognizes the spoken word and compares it with the target | Detects when a different word was said                    | Short child utterances can be misheard, does not judge pronunciation quality |
| Score calibration                         | Maps raw scores onto human ratings                         | Aligns scores with teacher judgement from small data      | Needs labelled examples from the target population                           |

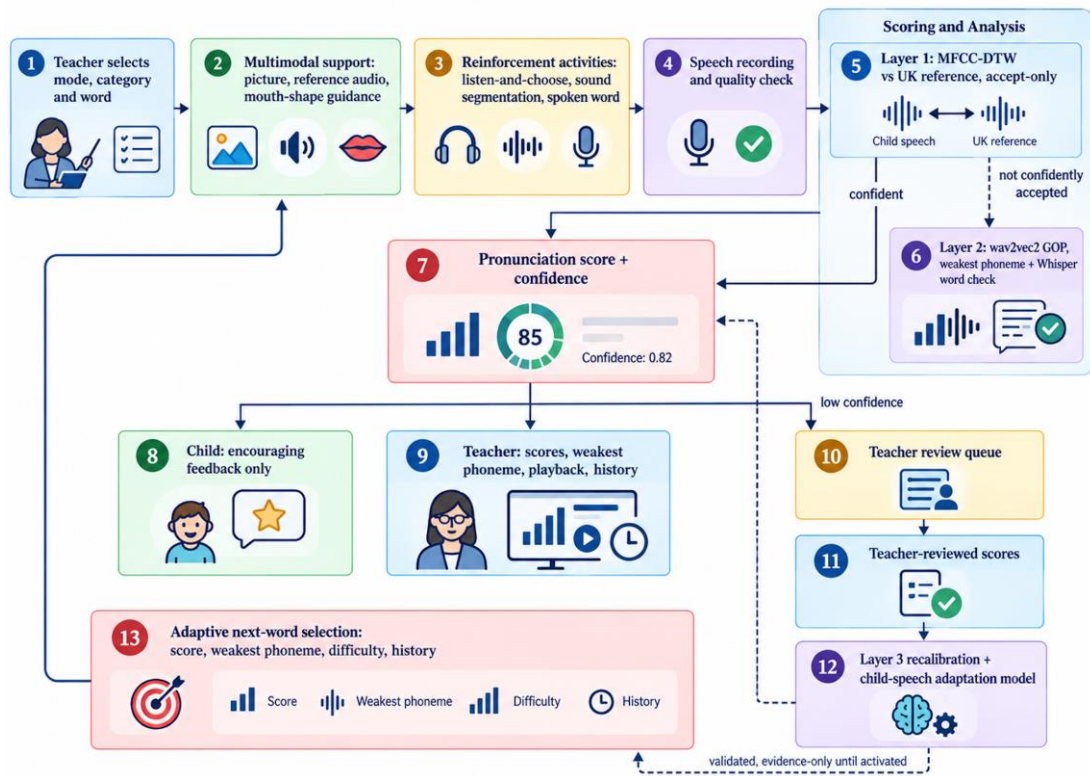
## 2.5 Gap synthesis and conceptual framework

Table 2.3: *How existing work covers the requirements of this component*

| Requirement                                         | ASD technology | CAPT and assessment | Children's speech | Existing apps | This component               |
|-----------------------------------------------------|----------------|---------------------|-------------------|---------------|------------------------------|
| Autism-sensitive, errorless interaction             | ✓              | ✗                   | ✗                 | Partial (AAC) | ✓                            |
| Pronunciation as the learning target                | ✗              | ✓                   | ✓                 | Partial       | ✓                            |
| Phoneme-level diagnosis                             | ✗              | ✓                   | Partial           | Partial       | ✓                            |
| Robust to children's voices                         | ✗              | ✗                   | Studied           | ✗             | ✓ (GOP-led, accept-only DTW) |
| Adaptation to recurring difficult sounds            | ✗              | Rare                | ✗                 | ✗             | ✓                            |
| Teacher review of uncertain scores                  | ✗              | ✗                   | ✗                 | ✗             | ✓                            |
| Learns from the target population's labelled speech | ✗              | Large corpora only  | Benchmarks only   | ✗             | ✓                            |
| Sri Lankan Grade 3 ESL context                      | ✗              | ✗                   | ✗                 | ✗             | ✓                            |

## Conceptual framework

*Figure 2.1: Conceptual framework of the Adaptive English Pronunciation Support Module*



## 3 RESEARCH METHODOLOGY

### 3.1 Research design

The study follows a **design and development research** approach, as set out in the proposal. In this approach an artefact is designed to solve a real-world problem, built, and evaluated, and the evaluation results feed back into the design. Here the artefact is the Adaptive English Pronunciation Support Module.

The research ran in four iterative phases:

1. **Analysis and content preparation.** Gathered requirements from the proposal's stakeholders (autistic learners, special education teachers), selected Grade 3 vocabulary, and prepared reference pronunciations, images and phoneme mappings.
2. **Design and development.** Built the learning interface, the backend services and the pronunciation scoring pipeline in increments, testing each before moving on.
3. **Scoring refinement.** The proposed word-level MFCC-DTW scoring was evaluated on real attempts from students. When it agreed poorly with teacher judgement, the design was extended with phoneme-level Goodness of Pronunciation scoring, word verification, teacher review, and recalibration. This phase is where the design changed in response to evidence.
4. **Evaluation and adaptation.** Compared system scores with teacher-reviewed scores and built a child-speech adaptation pipeline and model from classroom recordings of autistic students.

The evaluation is quantitative: it measures agreement between automated scores and teacher judgement. Teacher judgement is treated as the reference standard, as stated in the assumptions in Section 1.7.

### 3.2 Data participants or study materials

The component uses **78 target items**, each with a UK English reference pronunciation recorded by an adult speaker:

- the **26 letters** of the alphabet, and
- **52 words** in four categories aligned with the Grade 3 syllabus:
  - **animals**, e.g. cat, elephant, penguin
  - **fruits**, e.g. mango, papaya, guava
  - **classroom objects**, e.g. book, pencil, ruler
  - **everyday actions**, e.g. jump, eat, sleep.

For each word, a profile stores its phoneme sequence, syllable count, difficulty rating (1–5), consonant–vowel pattern, and lists of easier and related words used for adaptive selection.

Participants were students in special education settings, accessed through the partner school with guardian consent. Each student's population group is recorded in the system as entered by the school: ASD level 1, 2 or 3, or learning disability.

| Measure                                | Value                                                                      |
|----------------------------------------|----------------------------------------------------------------------------|
| Scored attempts                        | 245                                                                        |
| Students                               | 4                                                                          |
| Attempts by population                 | ASD level 1: 86, ASD level 2: 11, ASD level 3: 70, learning disability: 82 |
| Attempts flagged for teacher review    | 65                                                                         |
| Attempts with a teacher-reviewed score | 32                                                                         |

Each attempt stores the recording, all intermediate scoring evidence, and the teacher's score.

Ten classroom sessions were recorded in a teacher-prompt and child-repetition format: the teacher says a target word, and the child repeats it. The child also completes the

pronunciation activities for each section. The sessions cover animals, fruits, actions and letters, and teacher instructions are partly in Sinhala. These recordings are the raw material for the child-speech adaptation model.

### 3.3 Experimental or study procedure

1. **Setup.** The teacher logs in and selects the student. They choose alphabet or word mode, the category and the starting word.
2. **Support phase.** The child sees the picture, hears the reference pronunciation and views of activities related to each word. They can repeat this as often as needed.
3. **Reinforcement phase.** The child completes:
  - a. listen-and-choose
  - b. a sound-segmentation activity in which no speech is needed: sound chips are tapped in the order heard, and wrong taps get a gentle nudge rather than a "wrong" mark
  - c. spoken word recording.
4. **Recording.** Speech is recorded on the device as mono AAC (m4a) at 16 kHz and 64 kbit/s. The time between the prompt appearing and the start of recording is stored as a hesitation measure.
5. **Scoring and feedback.** The recording goes to the server for scoring. The child receives encouraging feedback only, while the teacher sees the detailed result.
6. **Next word.** The adaptive mechanism selects the next item.

Each recording passes through these stages:

1. **Quality checks.** The recording is rejected and the child asked to try again if any of the following holds:
  - a. total duration under 0.18 s,
  - b. voiced speech under 0.14 s,

- c. signal-to-noise ratio under 8 dB,
- d. more than 4% of samples clipped, or
- e. Peak amplitude under 0.015.

## 2. Layer 1: MFCC-DTW.

Audio is resampled to 16 kHz and pre-emphasized (coefficient 0.97).

It is framed into 512-sample frames with a 160-sample hop.

- a. Each frame becomes 13 MFCCs from a 26-filter mel filter bank, plus delta features.
- b. The result is aligned with the UK reference by DTW, and the normalized DTW distance is mapped to a 0–100 score by a sigmoid (midpoint 0.85, slope 0.18).
- c. If this score reaches **85**, the attempt is accepted. Otherwise, it goes to Layer 2.
- d. Layer 1 can **only accept** an attempt: a low DTW score never fails a child on its own, because DTW penalizes the difference between a child's voice and the adult reference.

## 3. Layer 2: Goodness of Pronunciation.

- a. The recording is padded with 0.3 s of silence on each side.
- b. It is passed through the pretrained Facebook/wav2vec2-lv-60-espeak-cv-ft phoneme recognizer [23].
- c. A GOP value [7] is computed for each target sound and mapped to a 0–100 score as  $100 \cdot \exp(0.55 \cdot \text{GOP})$ .
- d. The lowest-scoring sound is reported as the **weakest phoneme**.
- e. When GOP is available, the overall score combines GOP and DTW as  $0.7 \times \text{GOP} + 0.3 \times \text{DTW}$ .

- 4. **Word verification.** Whisper (base.en) [9], [28] transcribes the attempt. If it confidently hears a different vocabulary word, the attempt is rejected as a word

mismatch. If the speech is unclear, the attempt is marked unverified and flagged for teacher review rather than rejected.

5. **Confidence and review.** A confidence score is computed from the scoring evidence. Attempts with low confidence or unverified speech are flagged for teacher review.
6. **Layer 3: recalibration.** A recalibration curve fitted on teacher-reviewed scores for the student's population records what the calibrated score would be. It does not change the displayed score.

After each scored attempt, the next word is chosen as follows:

- **Remediation mode.** Used when the score is below 60, or when the weakest phoneme has been weak in at least two earlier attempts. Candidates come mainly from the current word's easier words.
- **Reinforcement mode.** Used otherwise. Candidates come mainly from related words.
- **Ranking.** Candidates are ranked by whether they contain the weak sound, difficulty compared with the current word, and the child's history.
- **Fallback.** If no candidate qualifies, the system selects another word containing the child's most frequently weak phoneme or continues the planned sequence.

Separately, when the same sound has been weakest in two or more saved attempts, a listen-and-choose round on that sound is added before the next speaking attempt.

1. **Segmentation.** Each session recording is converted to 16 kHz mono. Speech segments are found from frame energy, and a median pitch (F0) is estimated for each segment by autocorrelation.
2. **Speaker attribution.** Within each recording, segments are split into two pitch clusters by 2-means clustering on log F0:
  - a. the adult cluster, about 150–240 Hz,

- b. the child cluster, about 300–400 Hz.

Child segments are cut into separate clips, padded with background noise from the same room.

3. **Target alignment.** Each child clip takes its target word from the teacher prompt immediately before it, transcribed by Whisper with a vocabulary prompt. If that prompt cannot be identified, the target is left for the teacher to supply.
4. **Blind labelling.** Using an offline labelling tool, a teacher listens to each clip and records:
  - a. the target word,
  - b. a score from 0 to 100,
  - c. correct or incorrect,
  - d. whether to exclude the clip.

The system's score is hidden so that labels are independent of it.

5. **Feature extraction.** Each labelled clip goes through the production scoring pipeline. The following measurements are recorded as features:
  - a. Layer-1 accuracy and DTW distance,
  - b. GOP overall, per-sound mean, minimum and spread,
  - c. word verification status,
  - d. audio quality measures (SNR, voiced duration, silence ratio, speech onset),
  - e. word difficulty.
6. **Model training.** Two candidate models are trained to predict the teacher's score:
  - a. isotonic regression on the current overall score
  - b. ridge regression over all features

### 3.4 Measures and evaluation metrics

Table 3.2: *Evaluation metrics*

| Metric                    | Definition                                                                                                      | Purpose                                                                    |
|---------------------------|-----------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------|
| Mean absolute error (MAE) | Mean of $ \text{system score} - \text{teacher score} $                                                          | Average size of scoring error, in points out of 100                        |
| Pearson correlation (r)   | Linear correlation between system and teacher scores                                                            | Whether the system ranks attempts as the teacher does                      |
| Within-10 agreement       | % of attempts within 10 points of the teacher's score                                                           | Practical agreement                                                        |
| Pass/fail agreement       | % of attempts where system and teacher agree on reaching 70                                                     | Agreement on the decision a teacher acts on                                |
| Layer-1 accept rate       | % of attempts accepted by DTW alone                                                                             | Whether template matching works for child speech                           |
| Layer-1 correlation       | r between DTW accuracy (and distance) and teacher score                                                         | Answers RQ1                                                                |
| Review flag rate          | % of attempts flagged for teacher review                                                                        | How much work goes to the teacher                                          |
| Rejection rate            | % of attempts rejected by the quality or word-mismatch checks, and how many of those the teacher judged correct | Whether rejections are justified                                           |
| Held-out MAE improvement  | Baseline MAE minus candidate MAE on held-out students                                                           | Adoption rule for recalibration and the adaptation model ( $\geq 1$ point) |
| Functional correctness    | Pass rate of the automated test suite                                                                           | Software correctness                                                       |

### 3.5 Data analysis

For every attempt with a teacher-reviewed score, the displayed overall score is compared with the teacher's score using MAE, Pearson r, within-10 agreement and pass/fail agreement at 70. All analysis queries are read-only and leave out the stored audio.

A dedicated report computes:

- the Layer-1 accept rate,
- distributions of DTW accuracy and distance, overall and by population,
- correlation of each with teacher scores,
- a sweep of the accept threshold from 70 to 95, giving at each threshold the false-accept rate and the share of attempts escalated without need.

This answers RQ1: whether word-level template matching can assess these children.

For each population with at least **24** reviewed attempts:

- linear and isotonic recalibration curves are fitted and compared by **5-fold cross-validation** [19],
- a curve is accepted only if it reduces held-out MAE by at least **1 point**,
- linear slopes are limited to  $[0.25, 4]$ ,
- an isotonic fit must have at least 3 steps. A fit that collapses to a near-constant prediction shows there is no relationship to learn, so it is rejected.

Curves are fitted both for the adaptive score and for the displayed overall score. Because the review queue gives priority to uncertain attempts, the report also compares the mean confidence of reviewed and unreviewed attempts to show how biased the labelled set is.

- **Held-out evaluation.** Models are evaluated by grouped cross-validation (up to 5 folds). All clips from one session are held out together, so a model is always tested on a child it never saw in training.
- **Baseline.** The current overall score, without any adaptation.
- **Comparison set.** Rejected clips receive no score from the current system, so the score comparison uses scored clips only. Rejections are analysed separately against the teacher's correct/incorrect labels.

- **Adoption rules.** A model is adopted only if it beats the baseline's held-out MAE by at least 1 point, with at least 30 labelled clips from at least 3 sessions.
- **Interpretation.** Ridge [24] coefficients on standardized features are reported, showing which measurements best predict teacher judgement.

### 3.6 Ethics privacy security and responsible AI

#### Consent and participation

As set out in the ethical considerations:

- Participation was voluntary.
- Informed consent was obtained from parents or guardians before any child took part, and child assent was considered in a simple, understandable way.
- Children could withdraw at any time without consequence.
- Teachers could stop a session if a child became uncomfortable.

The component is an educational support tool, and its outputs are not interpreted as medical or diagnostic judgements.

#### Child-centered design

- Children never see numerical scores, and feedback is encouraging only.
- The sound-segmentation activity follows errorless-learning principles and works on days when a child doesn't speak.
- A reduced-stimulation setting turns off vibration and nudge animations.

#### Privacy and data protection

- **Pseudonymized identities.** Research outputs report only aggregate figures, so no student is identifiable.

- **Recordings in the database.** Recordings are stored in PostgreSQL and can only be accessed by authorized teachers. Analysis queries leave out the audio.
- **Retention.** A retention script deletes stored audio after a set period (60 days by default) while keeping the score records.
- **Classroom recordings.** These are kept outside the source code repository and outside cloud-synchronized storage and are excluded from version control. Clips are labelled in an offline tool, so children's voices are never uploaded to third-party services.
- **School name.** The partner school is not named in research outputs.
- **All speech processing is local.** The wav2vec2 model, Whisper and ffmpeg run on the project's own server, so no recordings are sent to external speech APIs.

## Security

- **Authentication:** JSON Web Tokens, with passwords hashed using bcrypt.
- **Access control:** by role (principal or teacher).
- **HTTP hardening:** security headers set with Helmet.
- **Rate limiting:** applies to authentication, to the API in general and, separately, to pronunciation scoring.
- **Input validation:** request bodies are validated, and validation errors never repeat submitted audio back in the response.

## Responsible AI

- **Human oversight.** Uncertain scores go to the teacher rather than being presented as certain, and the teacher remains the final judge.
- **No automatic failure.** The DTW layer cannot fail a child on its own, because it is known to penalise children's voices [17], [11].

- **Evidence-gated adaptation.** Recalibration and the adaptation model run in evidence-only mode and are adopted only after held-out validation shows they improve agreement with teachers.
- **Population-aware evaluation.** Calibration and reporting are done separately for each population, so results from one group are not assumed to transfer to another.
- **Known biases are reported.** The adult-reference mismatch, the non-random review sample and the pretrained model's lack of child-speech data are stated as limitations.

### 3.7 Reproducibility and research artefacts

Table 3.3: *Research artefacts*

| Artefact                      | Location                                                                                                                        |
|-------------------------------|---------------------------------------------------------------------------------------------------------------------------------|
| Tablet application            | auriva-frontend (React Native 0.81, Expo SDK 54)                                                                                |
| Backend and scoring services  | auriva-backend/src/services (Node.js, Express 5, Sequelize, PostgreSQL)                                                         |
| GOP worker                    | auriva-backend/ml/phoneme_gop_worker.py (Python 3.13, PyTorch 2.13, Transformers 5.14)                                          |
| Reference phoneme audit       | auriva-backend/ml/audit_reference_phonemes.py                                                                                   |
| Layer-1 and Layer-3 reports   | npm run layer1-report, npm run layer3-report                                                                                    |
| Adaptation pipeline and model | auriva-backend/ml/student_data/ (segmentation, attribution, labelling tool, feature extraction, training with scikit-learn 1.9) |
| Automated tests               | auriva-backend/test (87 tests), auriva-frontend/testing                                                                         |

Reproducibility measures:

- **Fixed models.** The phoneme recogniser (facebook/wav2vec2-lv-60-espeak-cv-ft) and the Whisper model (ggml-base.en) are fixed versions loaded from local copies, and the phoneme model loads offline.

- **Parameters in code.** All scoring parameters and adoption rules (Section 3.5) are named constants in the source code.
- **Deterministic analysis.** The analysis reports are deterministic, read-only scripts over the database. The training script fixes the evaluation procedure (grouped folds, adoption threshold) and writes both a machine-readable and a human-readable report.
- **Identical features.** Training and live scoring use the same feature extraction code, and the backend's prediction matches the training library's to within 0.0001 points.
- **Raw data not shared.** Children's recordings cannot be published under the consent terms. Code, parameters, aggregate results and trained model coefficients, which contain no audio, are available for review.

## 4 PRODUCT AND SOLUTION DESIGN

### 4.1 Requirements

#### 4.1.1 Stakeholders

| Stakeholder                 | Role in the component                                                                 |
|-----------------------------|---------------------------------------------------------------------------------------|
| Autistic children (Grade 3) | Multimodal practice, reinforcement activities, speaking attempts                      |
| Special education teachers  | Primary users: Run sessions, choose content, monitor results, review uncertain scores |
| School principals           | Institutional stakeholders supporting adoption. No direct use of the module           |
| Parents / guardians         | Give consent. Secondary stakeholders (progress view is future work)                   |

#### 4.1.2 Functional requirements

Table 4.1: Functional requirements and how they were met

| ID  | Requirement (from the proposal)                                   | Status        | How it is met                                                                              |
|-----|-------------------------------------------------------------------|---------------|--------------------------------------------------------------------------------------------|
| FR1 | Choice between alphabet and word pronunciation                    | Adapted       | The teacher chooses the mode at session setup, so the child doesn't face a decision screen |
| FR2 | Alphabet mode as an optional pathway without adaptive progression | Met           | Letters follow the planned sequence. The listen-and-choose round is skipped                |
| FR3 | Present the target with multimodal support                        | Met           | Picture, reference audio, phoneme breakdown and mouth-shape guidance                       |
| FR4 | Repeat the supports before reinforcement                          | Met           | Supports can be replayed as often as needed                                                |
| FR5 | Two to three reinforcement activities per word                    | Met           | Listen-and-choose, Tap the Sounds, spoken word                                             |
| FR6 | Capture speech during pronunciation activities                    | Met           | On-device recording (AAC, 16 kHz, mono)                                                    |
| FR7 | Preprocess speech before analysis                                 | Met           | Decoding, resampling, quality check, silence padding                                       |
| FR8 | Compare with the stored reference pronunciation                   | Met           | MFCC-DTW against the UK reference (Layer 1)                                                |
| FR9 | Produce a similarity-based score                                  | Met, extended | Word-level score plus phoneme-level GOP scores and the weakest phoneme                     |

|                 |                                                |         |                                                                                                                               |
|-----------------|------------------------------------------------|---------|-------------------------------------------------------------------------------------------------------------------------------|
| FR10            | Simple, child-friendly feedback                | Met     | Visual cues, celebration audio, encouraging messages                                                                          |
| FR11            | No detailed scores shown to the child          | Met     | Scores and evidence appear only in teacher views                                                                              |
| FR12            | Next word chosen from performance              | Met     | Adaptive next-word selection                                                                                                  |
| FR13            | Easier related word when below threshold       | Met     | Remediation mode when the score is below 60 or a weak sound recurs                                                            |
| FR14            | Planned progression when at or above threshold | Met     | Reinforcement mode, falling back to the planned sequence                                                                      |
| FR15            | Teacher selects the starting category          | Met     | Animals, classroom objects, fruits, daily actions                                                                             |
| FR16            | Teacher selects the starting word              | Met     | Word selection screen                                                                                                         |
| FR17            | Teacher views word-level scores                | Met     | Result screen and results history                                                                                             |
| FR18            | Category-wise performance percentages          | Adapted | Results history summarizes average score, trend, mastered sessions and recurring weak sounds. Category charts are future work |
| FR19            | Monitor progress over time                     | Met     | Trend indicator and session history                                                                                           |
| FR20            | Practice frequency and completion              | Met     | Mastered sessions out of total, and sessions with audio captured                                                              |
| FR21            | Replay the child's recordings                  | Met     | Replay in results history and the review queue                                                                                |
| FR22            | Keep per-student records                       | Met     | Each attempt stored with full scoring evidence                                                                                |
| FR23            | Identify repeated difficulties                 | Met     | "Recurring weak sounds" summary. A targeted listening round when a weak sound recurs                                          |
| FR24            | Secure teacher login                           | Met     | JWT authentication, teacher-role guard                                                                                        |
| FR25<br>(added) | Teacher review of uncertain scores             | Added   | Review queue ranked by uncertainty and population coverage                                                                    |
| FR26<br>(added) | Check that the target word was spoken          | Added   | Whisper word verification                                                                                                     |
| FR27<br>(added) | Learn from teacher judgement                   | Added   | Layer 3 recalibration and the child-speech adaptation model (evidence-only until validated)                                   |

#### 4.1.3 Non-functional requirements

Table 4.2: Non-functional requirements

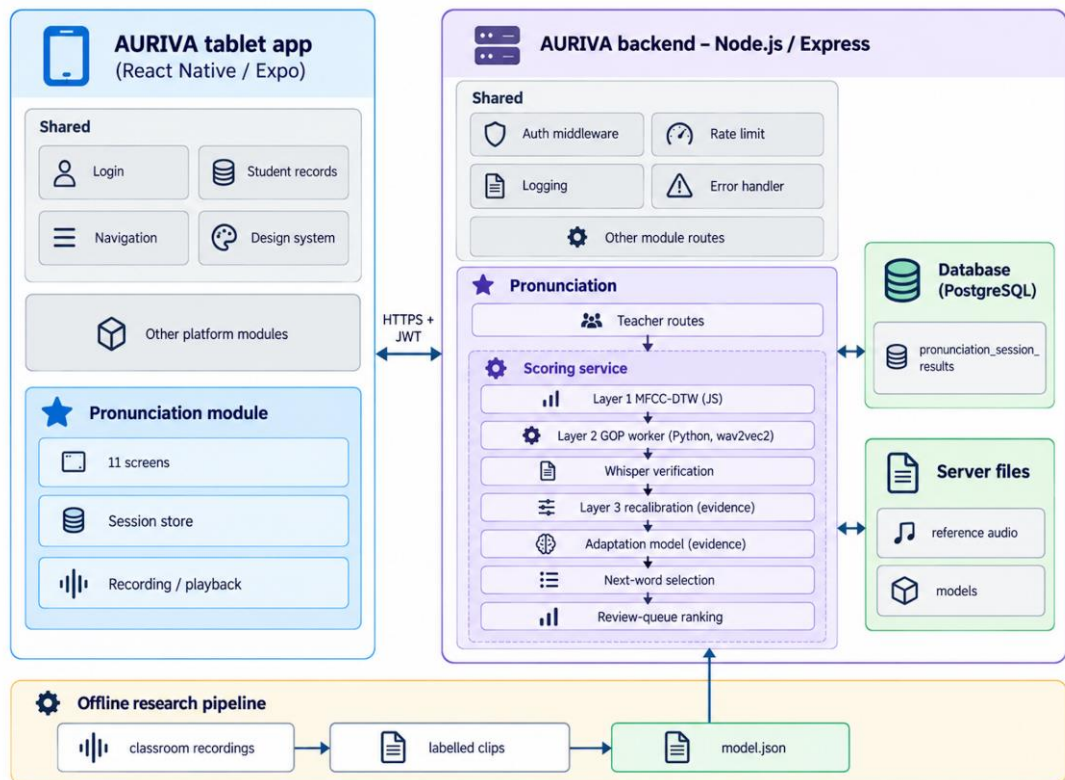
| ID | Requirement (from the proposal) | Design response |
|----|---------------------------------|-----------------|
|----|---------------------------------|-----------------|

|       |                                                        |                                                                                                                               |
|-------|--------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------|
| NFR1  | Simple, clear interface for children and teachers      | One task per screen, large targets, consistent layout, the student's avatar theme                                             |
| NFR2  | Acceptable response time                               | About 3 s per attempt with the model loaded. The GOP worker is kept running so the model loads only once                      |
| NFR3  | Store data accurately without loss                     | The server saves each scored attempt itself. The client completes it by reference. Only read requests are retried             |
| NFR4  | Consistent scores against a UK reference               | Fixed reference set, fixed model versions, named scoring constants                                                            |
| NFR5  | Only authorized teachers access records                | JWT, teacher-role guard, rate limits, validated inputs                                                                        |
| NFR6  | Easy to extend                                         | Word bank and word profiles are data. New words need a profile, a picture and reference audio                                 |
| NFR7  | Scale to more students, words and categories           | Per-population analysis. Reports read only score columns, never the audio                                                     |
| NFR8  | Works on tablets and phones                            | React Native with Expo for iOS and Android                                                                                    |
| NFR9  | Autism-friendly, minimal distraction                   | Errorless activities, reduced-stimulation option, no scores shown to children                                                 |
| NFR10 | Available during sessions. Teachers can review records | Server crash recovery with automatic restart. Fallback scoring if the GOP worker is down. History and review always available |

## 4.2 Component architecture within the complete solution

The module is one component of the AURIVA platform. It shares the platform's tablet app, authentication, student records and backend, and owns its own screens, services, data and machine-learning worker.

*Figure 4.2: The component within the complete solution*



Layers:

1. **Presentation (tablet app).** Eleven screens in the teacher navigation stack, a session store held in memory and on the device, and helpers for recording and playback.
2. **API.** Teacher routes protected by JWT [29] and a teacher-role guard. The scoring endpoint has its own rate limit (30 per 5 minutes per teacher).
3. **Services.**
  - a. scoring orchestration,
  - b. Layer 1 (MFCC-DTW, JavaScript),
  - c. Layer 2 (a Python GOP worker that stays running),
  - d. Whisper word verification,
  - e. Layer 3 recalibration,
  - f. the child-speech adaptation model,
  - g. adaptive next-word selection,

- h. review-queue ranking.
- 4. **Data.** PostgreSQL table `pronunciation_session_results`, with reference audio and model files stored on the server.
- 5. **Offline research pipeline.** Scripts that turn classroom recordings into labelled child clips and train the adaptation model. They are separate from the live server and write only a model file.

**Shared with the platform:** login and JWT, student records (ID, population group, avatar theme, reduced-stimulation setting), navigation, the backend server and its middleware, the database, and the design system.

### 4.3 Design alternatives and decisions

Table 4.3: Design alternatives and decisions

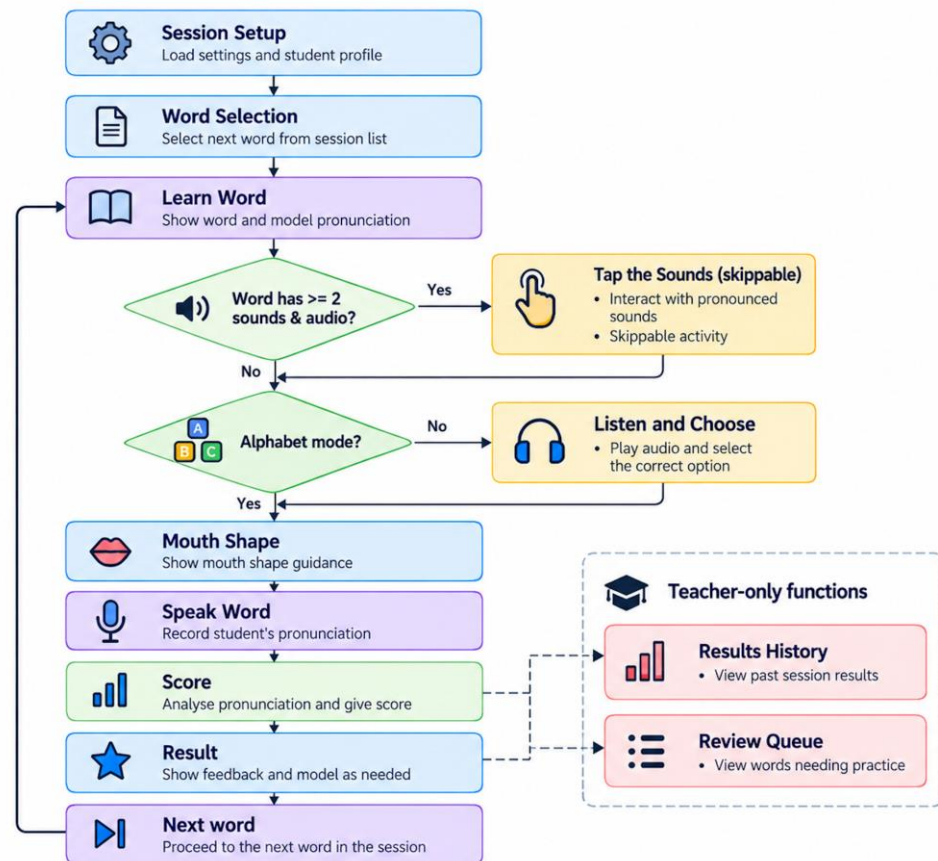
| Decision              | Alternatives considered                                                                  | Chosen                                | Reason                                                                                                                                                                                                                 |
|-----------------------|------------------------------------------------------------------------------------------|---------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Backend framework     | FastAPI (proposal) or Node.js/Express                                                    | Node.js/Express, plus a Python worker | Shares the platform's existing backend, authentication and database. Python used only for model inference                                                                                                              |
| Main scoring method   | MFCC-DTW only (proposal), end-to-end neural model, cloud pronunciation API, wav2vec2 GOP | Cascade: DTW then GOP                 | DTW needs no data but doesn't suit child voices. End-to-end models need corpora that don't exist. Cloud APIs send children's voices to third parties and cost money. GOP gives phoneme-level evidence without training |
| Role of DTW           | Pass/fail decision, blend only, accept-only check                                        | Accept-only check plus 30% blend      | A low DTW score must never fail a child, because children's voices differ from the adult reference                                                                                                                     |
| Word verification     | None, cloud speech recognition, local Whisper                                            | Local Whisper (base.en)               | Detects wrong-word attempts. Local processing keeps audio private                                                                                                                                                      |
| Running the GOP model | Start a process per request, keep one worker running                                     | One long-running worker               | The 1.2 GB model loads once. Crash cooldown and restart keep it available                                                                                                                                              |
| Where scoring runs    | On the device, on the server                                                             | Server                                | Models too large for phones. One consistent scoring version for all devices                                                                                                                                            |

|                               |                                                               |                                                      |                                                                                         |
|-------------------------------|---------------------------------------------------------------|------------------------------------------------------|-----------------------------------------------------------------------------------------|
| Who writes scores             | Client sends scores, server writes them                       | Server writes. Client completes the row by reference | Prevents tampering and duplicate rows                                                   |
| Feedback to the child         | Show scores, encouragement only                               | Encouragement only                                   | Errorless learning. Avoids discouraging children                                        |
| Applying recalibration        | Apply as soon as fitted, evidence-only until validated        | Evidence-only, gated by cross-validation             | A correction is applied only when it measurably helps                                   |
| Adaptation model              | Deep model, small linear or isotonic models                   | Ridge and isotonic                                   | Works with little data, explainable, runs in the backend as plain JSON                  |
| Visual pronunciation guidance | Lip-sync video (proposal), drawn mouth shapes                 | SVG mouth shapes per phoneme                         | Clear and calm. No video to produce or load. Easy to extend to every phoneme            |
| Teacher monitoring            | Separate web dashboard (proposal), teacher screens in the app | In-app teacher screens                               | One app for the whole session. Review happens where the teacher already is              |
| Audio storage                 | Separate blob storage, in the database                        | In the database, with a retention script             | Simple, transactional with the score row. Old audio can be removed while keeping scores |
| Development delivery          | Native development build only, Expo Go                        | Expo Go for iOS, plus a native Android build         | No iOS build setup was needed to test on an iPhone                                      |

## 4.4 Detailed design

### 4.4.1 Session flow

*Figure 4.4: Session flow*

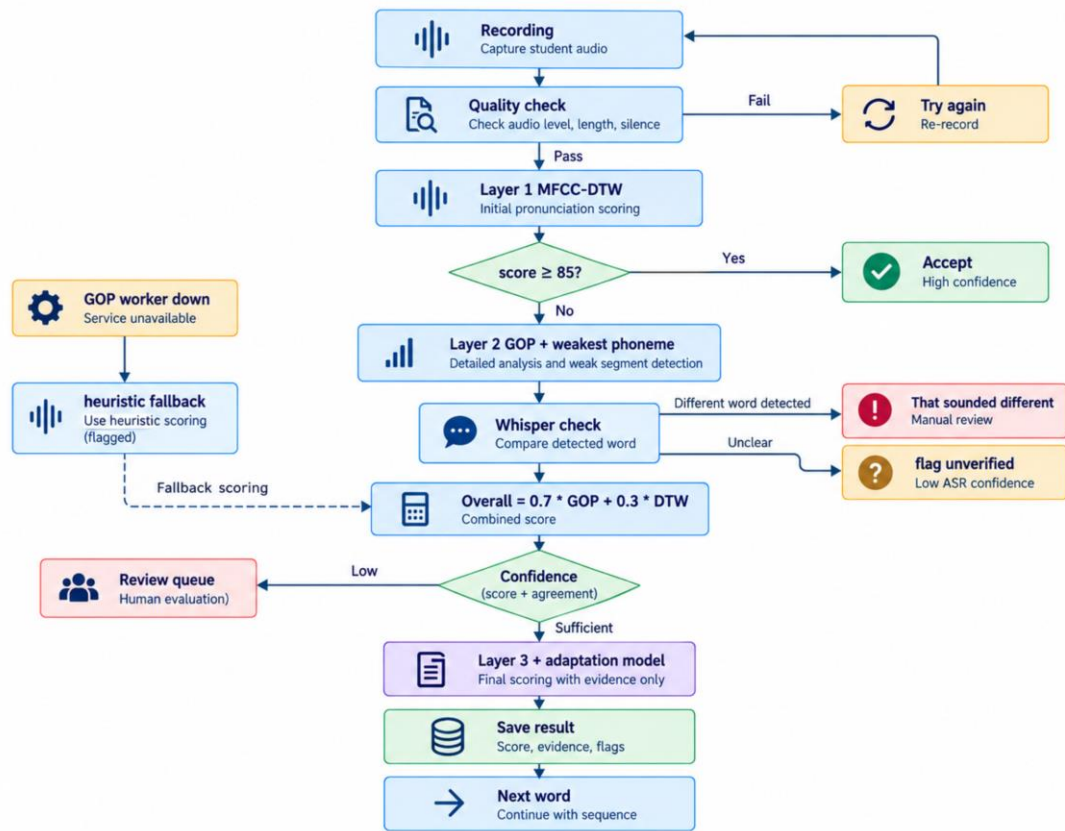


1. **Session Setup:** mode, category, reduced-stimulation setting.
2. **Word Selection.**
3. **Learn Word.**
4. **Tap the Sounds:** if the word has two or more sounds and audio is available.  
The teacher can skip it.
5. **Listen and Choose:** skipped in alphabet mode.
6. **Mouth Shape.**
7. **Speak Word:** record, then score.
8. **Result:** the child gets feedback, and the teacher sees score, weakest phoneme and recommendation.
9. **Next word:** returns to Learn Word.

Teacher-only screens: Results History and Review Queue.

## 4.4.2 Scoring design

Figure 4.5: Scoring cascade



### 1. Quality check. Reject if any of these holds:

- total duration under 0.18 s,
- voiced speech under 0.14 s,
- SNR under 8 dB,
- more than 4% of samples clipped,
- peak amplitude under 0.015.

A rejected attempt asks the child to try again.

### 2. Layer 1.

- Resample to 16 kHz and apply pre-emphasis (0.97).
- 512-sample frames, 160-sample hop, Hamming window.

- c. 26 mel filters, 13 MFCCs plus deltas.
- d. DTW alignment against the reference. The distance is mapped to a 0–100 score by a sigmoid (midpoint 0.85, slope 0.18).
- e. A score of 85 or more is accepted. Otherwise, the attempt goes to Layer 2.

### 3. Layer 2.

- a. Pad the audio with 0.3 s of silence on each side.
- b. wav2vec2 phoneme posteriors [23] give a GOP value [7] per target sound, mapped to a score as  $100 \cdot \exp(0.55 \cdot \text{GOP})$ .
- c. The lowest-scoring sound is the weakest phoneme.
- d. Overall score =  $0.7 \times \text{GOP} + 0.3 \times \text{DTW}$ .

### 4. Word verification (Whisper).

- a. Confidently a different vocabulary word: rejected as a word mismatch.
- b. Unclear speech: kept and flagged as unverified.

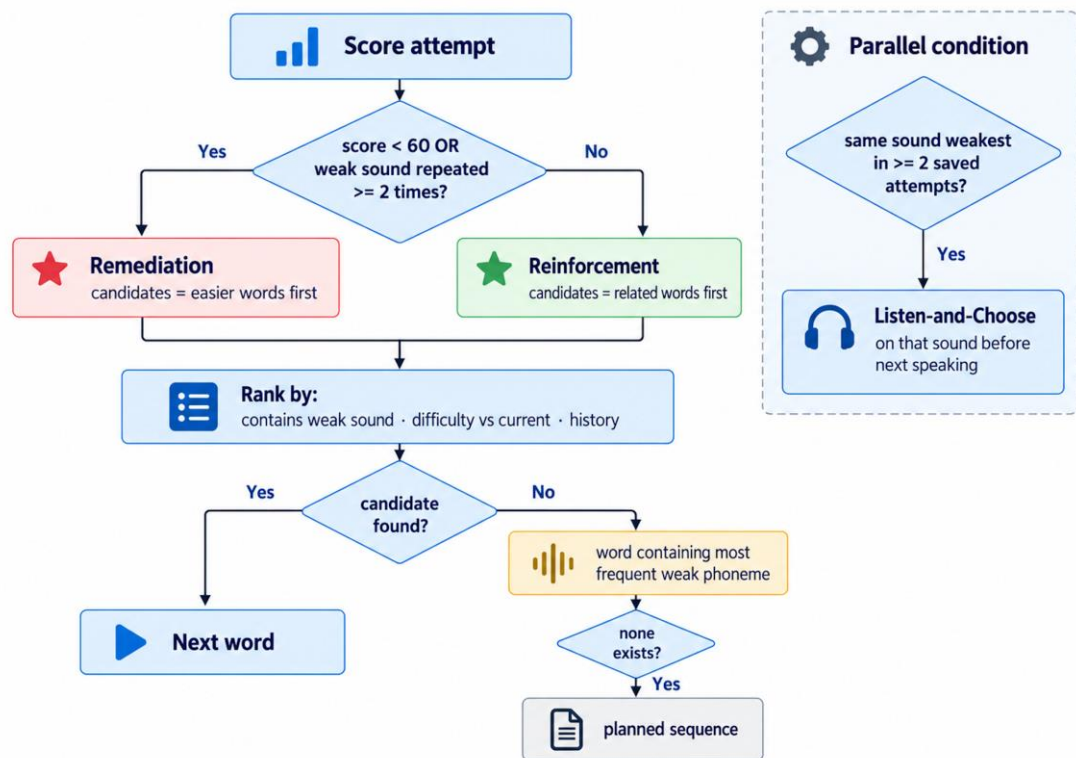
5. **Confidence and review flag.** Low confidence or unverified speech flags the attempt for teacher review.

6. **Layer 3 and the adaptation model.** Both record evidence only.

7. **Fallback.** If the GOP worker is unavailable, a heuristic score is used and the result is flagged.

#### 4.4.3 Adaptive next-word design

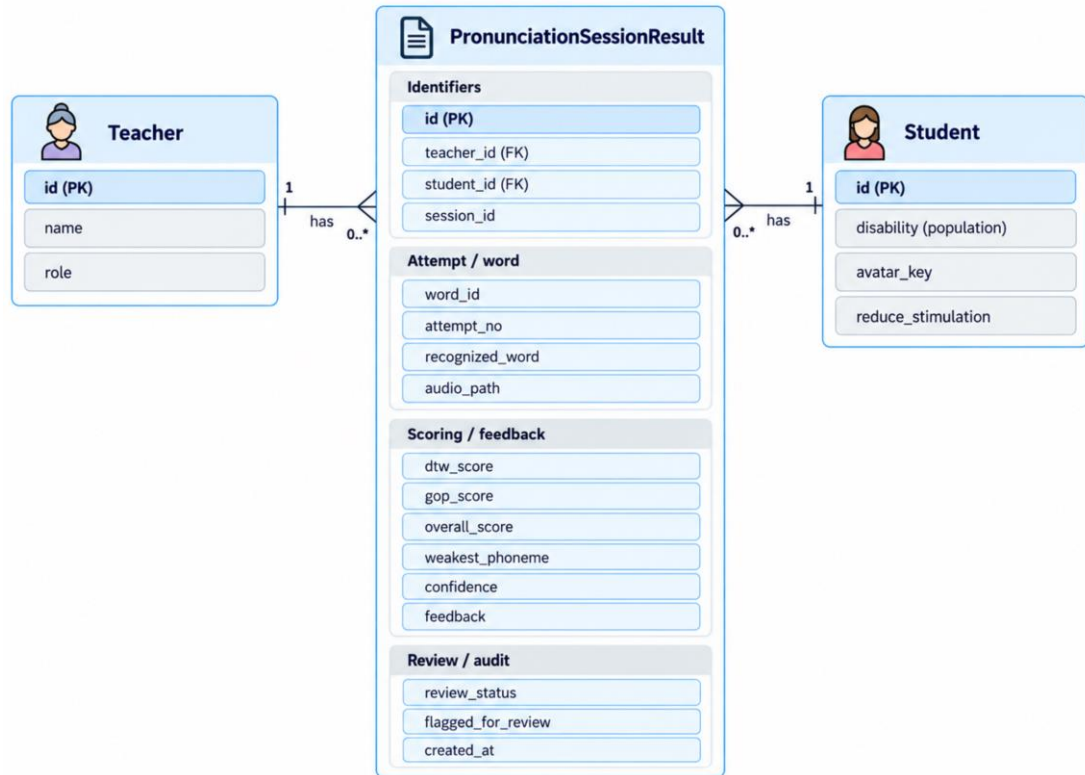
*Figure 4.6: Next-word decision*



- **Mode.** Remediation if the score is below 60 or the weakest phoneme has been weak in two or more earlier attempts, and reinforcement otherwise.
- **Candidates.** Remediation draws first on easier words, while reinforcement draws first on related words.
- **Ranking.** Whether the candidate contains the weak sound, its difficulty compared with the current word, and the child's history.
- **No candidate.** Another word containing the child's most frequent weak phoneme, otherwise the planned sequence.
- **Separate rule.** When the same sound has been weakest in two or more saved attempts, a listen-and-choose round on that sound comes before the next speaking attempt.

#### 4.4.4 Data design

Figure 4.7: Entity–relationship view



PronunciationSessionResult: one row per attempt.

| Group          | Fields                                                                                                                                                                |
|----------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Keys           | id, teacher_id (references Teacher), student_id (references Student)                                                                                                  |
| Target         | mode (word or alphabet), category_id, word_id, word_label                                                                                                             |
| Scores         | overall_score, adaptive_score, confidence_score, confidence_level, segmental_accuracy, dtw_distance, layer1_decision, scoring_method                                  |
| Evidence       | phoneme_scores (JSONB), speech_verification (JSONB), recognized_text, recommendation_details (JSONB, holding the full scoring evidence, Layer 3 and adaptation model) |
| Behaviour      | response_duration, hesitation_time, heard_reference_audio, listen_choose_data (JSONB)                                                                                 |
| Recommendation | recommendation_type, recommendation_message, next_word_id                                                                                                             |
| Review         | needs_teacher_review, teacher_reviewed_score (0–100), teacher reviewed at, teacher reviewed by                                                                        |
| Audio          | raw_audio_data (binary), raw_audio_mime_type, raw_audio_size                                                                                                          |
| Workflow       | workflow_completed, created_at                                                                                                                                        |

**Relationships:** a teacher has many results, a student has many results.

#### 4.4.5 API design

Six REST endpoints under /api/teacher, all JWT-protected:

| Method | Path | Purpose |
|--------|------|---------|
|--------|------|---------|

|       |                                         |                                 |
|-------|-----------------------------------------|---------------------------------|
| POST  | /students/:id/pronunciation-score       | Score an attempt                |
| POST  | /students/:id/pronunciation-results     | Save a completed session result |
| GET   | /students/:id/pronunciation-results     | Results history                 |
| GET   | /pronunciation-results/:resultId/audio  | Replay audio                    |
| GET   | /pronunciation-review-queue             | Review queue                    |
| PATCH | /pronunciation-results/:resultId/review | Save a teacher's review         |

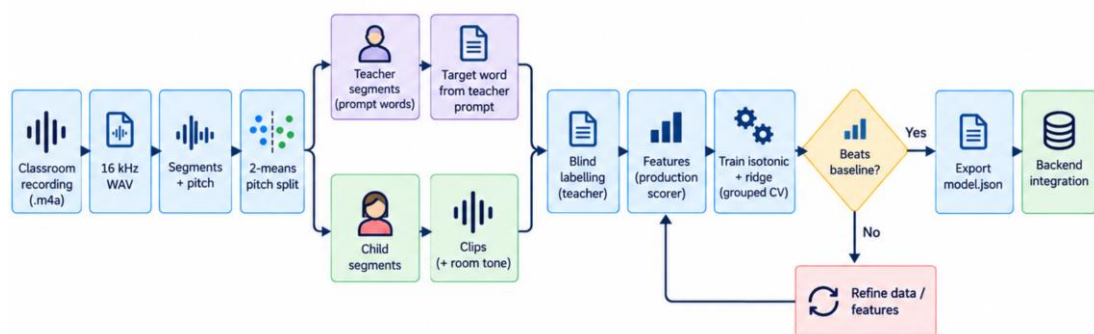
Request bodies are validated. Error responses use consistent codes (AUDIO\_QUALITY\_FAILED, WORD\_MISMATCH, DATABASE\_BUSY) so the app can show friendly messages.

#### 4.4.6 Interface design principles

- One task per screen, in a predictable order, with the same layout throughout.
- Pictures and audio before any text-heavy instruction.
- Encouragement only, never "wrong" marks, and wrong taps get a gentle nudge.
- A reduced-stimulation setting turns off praise voice prompts and nudge animations.
- The student's avatar theme colours the screens for familiarity.
- Teacher-only information (scores, confidence, evidence) is kept out of the child's view.

#### 4.4.7 Child-speech adaptation design

Figure 4.8: Adaptation pipeline



1. Classroom recording (m4a).
2. Convert to 16 kHz WAV.
3. Find speech segments and estimate pitch.
4. Separate teacher from child by 2-means pitch clustering.
5. Cut child clips, padded with the room's background noise.
6. Take each clip's target word from the teacher's prompt.
7. Blind labelling by a teacher.
8. Extract features with the production scorer.
9. Train isotonic and ridge models with grouped cross-validation.
10. If a model beats the baseline by at least 1 point, write model.json.
11. The backend records the model's prediction as evidence.

#### 4.5 Standards and professional practice

- **Requirements:** functional and non-functional requirements expressed as testable "the system shall" statements, following ISO/IEC/IEEE 29148 practice [30], and traced to the implementation (Tables 4.1 and 4.2).
- **Software quality:** non-functional requirements mapped to ISO/IEC 25010 [31] qualities: usability, performance efficiency, reliability, security, maintainability and portability.
- **API design:** RESTful resources, standard HTTP methods and status codes, and an OpenAPI (Swagger) specification served at /api-docs.
- **Security:**
  - JWT authentication (RFC 7519 [29]),
  - bcrypt password hashing,
  - role-based access,

- rate limiting,
- input validation,
- security headers (Helmet).

Together these follow OWASP API Security Top 10 practice.

- **Data protection:** consent, collecting only the data needed, purpose limitation, retention controls and restricted access, consistent with Sri Lanka's Personal Data Protection Act No. 9 of 2022 [32] and the ethics commitments in the proposal. No children's audio is sent to third-party services.
- **Accessibility and inclusive design:** informed by WCAG 2.1 [33] principles (perceivable, operable, understandable) and autism-friendly design: predictable flow, reduced stimulation, errorless feedback.
- **Engineering practice:**
  - Git version control,
  - automated tests (87 backend, 16 frontend pronunciation),
  - structured logging,
  - configuration through environment variables, keeping secrets out of the code,
  - documented, reproducible research scripts.
- **Research practice:**
  - IEEE citation style,
  - held-out validation before any model is adopted,
  - blind labelling for the adaptation data,
  - reporting of limitations.

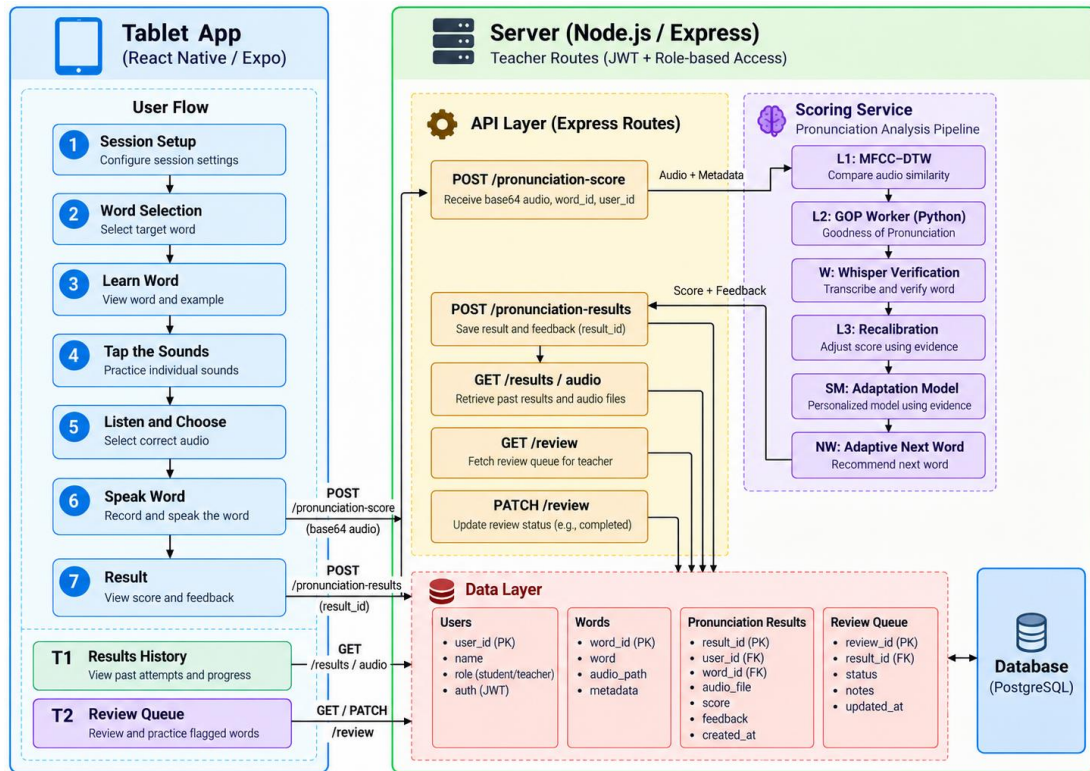
## 5 IMPLEMENTATION INTEGRATION AND TESTING

### 5.1 Development environment and technology stack

Table 5.1: Technology stack of the Pronunciation Support component

| Layer                | Technology                                                                                      | Role in the component                                                                 |
|----------------------|-------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------|
| Tablet application   | React Native 0.81, Expo SDK 54, React 19                                                        | Cross-platform learning and teacher interface (iOS and Android)                       |
| Navigation and state | React Navigation 7 (native stack), Zustand, AsyncStorage                                        | Screen flow, session state that survives an app restart                               |
| Audio on device      | expo-av, expo-asset, expo-file-system                                                           | Recording (m4a/AAC, 16 kHz, mono), reference audio playback, replay of saved attempts |
| Graphics             | react-native-svg, expo-linear-gradient                                                          | Mouth-shape illustrations, themed backgrounds                                         |
| Backend API          | Node.js, Express 5                                                                              | Scoring, results and review endpoints                                                 |
| Database             | PostgreSQL (Azure), Sequelize ORM                                                               | Attempts, scores, evidence, teacher reviews, audio                                    |
| Signal processing    | Custom MFCC and DTW in JavaScript (fft.js), ffmpeg                                              | Layer-1 features and alignment, audio decoding and resampling                         |
| Phoneme scoring      | Python 3.13, PyTorch 2.13, Hugging Face Transformers 5.14, facebook/wav2vec2-lv-60-espeak-cv-ft | Layer-2 Goodness of Pronunciation worker                                              |
| Word verification    | whisper.cpp (whisper-cli) [28], ggml-base.en model                                              | Checks the target word was spoken                                                     |
| Adaptation model     | scikit-learn 1.9, NumPy, SoundFile                                                              | Training the child-speech adaptation model                                            |
| Security             | jsonwebtoken, bcryptjs, Helmet, express-rate-limit, express-validator                           | Authentication, hardening, rate limiting, input validation                            |
| API documentation    | swagger-jsdoc, swagger-ui-express                                                               | Interactive API documentation at /api-docs                                            |
| Testing              | Node.js test runner (node --test), Jest with jest-expo, React Native Testing Library            | Backend and frontend automated tests                                                  |
| Tools                | Visual Studio Code, Git, Postman, Expo Go, Android emulator                                     | Development, version control, API testing, device testing                             |

## 5.2 Component implementation



## 5.3 Integration with the other components

The component plugs into the shared platform at these points only:

- 1. Authentication and roles.** Pronunciation endpoints sit on the shared teacher router (router.use(verifyToken, isTeacher)) and use the platform's login and JWT tokens. The tablet app's shared API client attaches the token to every request.
- 2. Student records.** Each session belongs to a student from the platform's student records, identified by student ID. The component reads the student's population group (the disability field), which drives the review queue's coverage boost and the per-population calibration, and the avatar theme used to style the screens.
- 3. Navigation.** The pronunciation screens are registered in the teacher navigation stack and open from the student's detail screen. The review queue is reached from the teacher's area.

4. **Shared backend and database.** The component runs inside the platform's Express server and database. It registers its own model (PronunciationSessionResult), routes, validation and services, and uses the server's existing middleware (Helmet, rate limiting, logging, error handling). Its database table is added by the server's existing start-up synchronization.
5. **Shared design system.** Colors, layout constants, toast notifications, and common components are reused, so the module looks like the rest of the app.

**Integration testing.** Integration was verified end to end on a physical Android tablet (Expo Go) and an Android emulator against the running backend. The test covered login, student selection, a full word session through recording, scoring, the result screen and saving, history replay, and submitting a review from the queue.

## 5.4 Verification and software testing

### Automated tests

*Table 5.2: Automated tests for the component*

| Suite                                  | What it verifies                                                                                                                                                                                                                                   |
|----------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| layer1Cascade.test.js                  | DTW gives zero distance for identical sequences and growing distance as sequences diverge. Distance-to-score mapping decreases and stays within 0–100. The Layer-1 gate is accept-only. The quality check rejects silence and accepts a clear tone |
| speechVerificationGate.test.js         | Transcript normalization. Exact and near word matches. Spoken letter names. Different short letters never matched by similarity. Garbled speech not treated as a confident mismatch                                                                |
| pronunciationScoringService.test.js    | Adaptive model recommends continue, remediate, or lower confidence for poor audio. Fallback when audio or the word is missing. Weak-phoneme history counts only real acoustic attempts                                                             |
| alphabetPronunciation.test.js          | All 26 letters have profiles and reference audio. Every letter's reference recording completes the full scoring flow                                                                                                                               |
| adaptiveCalibration.test.js (20 tests) | Linear and isotonic fitting, cross-validation adoption rule, slope limits, rejection of collapsed isotonic fits, population tags, fallback when too few labels, compact stored output                                                              |

|                                                                        |                                                                                                                    |
|------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------|
| pronunciationReviewQueue.test.js                                       | Informativeness ranking (low confidence first, boost for under-covered populations, bounds)                        |
| teacherPronunciationHistory.test.js                                    | History order. Saved results report only audio that was actually stored                                            |
| studentAdaptationModel.test.js                                         | Ridge and isotonic prediction, filling in missing features, feature extraction, no output when no model is trained |
| moduleLoad.test.js                                                     | All services load without errors                                                                                   |
| Frontend pronunciation-history.test.js, pronunciation-payloads.test.js | History formatting. The scoring request payload sent by the app                                                    |

### Pipeline verification

- **Reference audit.** `audit_reference_phonemes.py` runs every reference recording through the GOP engine. A reference that scores low on its own word shows a phoneme-mapping gap or a poor recording. The audit led to:
  - alternative tokens for consonant clusters (e.g. *buffalo* rose from 13 to 100),
  - 0.3 s silence padding before inference,
  - accent variants of several vowels and consonants.

It also flagged reference recordings that need re-recording.

- **Integration smoke tests.** `verifyLayer1Layer2Integration.js` and `fullCorpusSmokeTest.js` run the full cascade on the reference set.
- The backend's adaptation model predictions match scikit-learn's to within 0.0001 points.

Table 5.3: Defects found and fixed during verification

| Defect                                               | Cause                                                         | Fix                                                                            |
|------------------------------------------------------|---------------------------------------------------------------|--------------------------------------------------------------------------------|
| Saved recordings would not play back on iOS          | iOS playback cannot open data: URIs                           | Clips are written to a cache file and played from there                        |
| Double-tapping the microphone gave a recording error | A second recorder started while the first was still preparing | A start guard. The microphone is released if the screen closes while preparing |

|                                                 |                                                                                   |                                                                  |
|-------------------------------------------------|-----------------------------------------------------------------------------------|------------------------------------------------------------------|
| Scoring hung on WAV input                       | ffmpeg's input and output were the same path, so it waited at an overwrite prompt | A separate output file and the -y flag                           |
| GOP worker failed to start                      | The Python environment pointed to an uninstalled interpreter                      | Pointed the environment at a supported Python 3.13               |
| The adaptation model read missing features as 0 | Number(null) evaluates to 0                                                       | Missing values are explicitly filled in from training statistics |

### Device testing

The component was tested in Expo Go on a physical tablet and on an Android emulator. Testing covered the recording quality check, the microphone permission prompt, audio playback in silent mode, recovery from network errors, and resuming an interrupted session.

## 5.5 Deployment and operation

- **Tablet app.**
  - **Development and pilot:** the app is served by Expo's Metro bundler over the local network and opened in Expo Go (npm start, which uses Expo Go mode).
  - **Android:** a native development build is available (npm run android).
  - **Backend address:** the app works out the backend address from the address it was loaded from, so no manual configuration is needed on the classroom network.
- **Backend.**
  - Runs as a Node.js service (npm start) on the project server.
  - Connects to an Azure PostgreSQL database.
  - On start-up, it synchronizes the database schema and starts the GOP worker in the background.
  - All speech models run locally on the server, so no recordings are sent to external services.

- **Configuration.** Environment variables control the database connection, JWT secret, CORS, rate limits, the GOP worker (Python path, timeouts, on/off switch), Whisper (binary path, model path, threads), the reference audio folder, and the adaptation model path.
- **Operation and monitoring.**
  - Logs are structured (Winston).
  - Process-level handlers catch unhandled errors.
  - Read-only reports (`npm run layer1-report`, `npm run layer3-report`) monitor scoring quality and calibration readiness.
  - `npm run purge-old audio` removes stored audio older than the retention period while keeping the score records.
- **Hardware.** The GOP worker needs about 1.3 GB of memory and runs on an 8 GB development machine.

## 6 RESULTS AND VALIDATION

### 6.1 Research results

#### Dataset overview

From 7 May to 9 October 2026, the component recorded **249 scored attempts from 5 students**. Of these, **65 (26.1%)** were flagged for teacher review and **32** received a teacher-reviewed score.

Table 6.1: *Attempts and teacher reviews by population*

| Population          | Attempts | Teacher-reviewed |
|---------------------|----------|------------------|
| ASD level 1         | 86       | 2                |
| ASD level 2         | 11       | 3                |
| ASD level 3         | 70       | 2                |
| Learning disability | 82       | 25               |
| Total               | 249      | 32               |

#### Agreement between system scores and teacher judgement

For the 32 teacher-reviewed attempts, the overall score shown to teachers was compared with the teacher's score.

Table 6.2: *Agreement between the system's overall score and teacher scores (n = 32)*

| Metric                       | Value                    |
|------------------------------|--------------------------|
| Pearson correlation (r)      | 0.73 (95% CI 0.51–0.86)  |
| Mean absolute error          | 7.9 points (0–100 scale) |
| Within 10 points of teacher  | 71.9%                    |
| Agreement on pass/fail at 70 | 87.5%                    |

The system's scores follow the teacher's judgement closely. On average they differ by under 8 points, and in seven of eight cases the system and the teacher agree on whether the attempt reaches the 70-point threshold.

### Layer 1: word-level MFCC-DTW against an adult reference (RQ1)

151 attempts recorded a Layer-1 decision.

Table 6.3: *Layer-1 outcomes*

| Measure                                                | Value                   |
|--------------------------------------------------------|-------------------------|
| Attempts accepted by DTW alone (accuracy $\geq 85$ )   | 0 of 151 (0%)           |
| Attempts escalated to Layer 2                          | 151 (100%)              |
| Layer-1 accuracy, median (10th–90th percentile)        | 72 (64–75), range 50–81 |
| Normalized DTW distance, median (10th–90th percentile) | 0.681 (0.645–0.743)     |

Layer-1 accuracy fell in a narrow band whatever the actual quality of the pronunciation: 95% of attempts scored between 60 and 80. No population came close to the accept threshold. Mean accuracy was 70.8 for ASD level 1, 69.3 for ASD level 2, 66.0 for ASD level 3 and 70.2 for learning disability.

Table 6.4: *Layer-1 agreement with teacher scores*

| Measure                                          | Value                       |
|--------------------------------------------------|-----------------------------|
| r, Layer-1 accuracy vs teacher (n = 30)          | 0.19 (95% CI –0.18 to 0.52) |
| r, DTW distance vs teacher (n = 31)              | –0.24                       |
| Mean absolute error, Layer-1 accuracy vs teacher | 19.9 points (bias –2.8)     |

The confidence interval for the Layer-1 correlation includes zero, so word-level DTW against an adult reference showed **no reliable relationship** with teacher judgement. Its error (19.9 points) is more than double that of the full pipeline (7.9 points).

Table 6.5: *Sweep of the Layer-1 accept threshold (teacher-reviewed attempts, where "correct" means a teacher score of at least 70)*

| Threshold | Accepted | False accepts | Escalated | Escalated needlessly |
|-----------|----------|---------------|-----------|----------------------|
| 70        | 17       | 17.6%         | 13        | 61.5%                |
| 75        | 4        | 25.0%         | 26        | 73.1%                |
| 80        | 1        | 0.0%          | 29        | 72.4%                |

| Threshold | Accepted | False accepts | Escalated | Escalated needlessly |
|-----------|----------|---------------|-----------|----------------------|
| 85 (used) | 0        | —             | 30        | 73.3%                |
| 90        | 0        | —             | 30        | 73.3%                |

No threshold works. A lower threshold lets DTW accept attempts, but between 17.6% and 25% of those acceptances are wrong. A higher threshold means DTW never decides anything. Answer to RQ1: word-level MFCC-DTW against an adult UK reference cannot reliably assess the pronunciation of these children.

## Layer 2: phoneme-level GOP (RQ2)

Because Layer 1 accepted nothing, every scored attempt was decided mainly by Layer 2: the overall score is  $0.7 \times \text{GOP} + 0.3 \times \text{DTW}$ . The agreement in Table 6.2 ( $r = 0.73$ , mean absolute error 7.9) therefore reflects the GOP-based pipeline. Compared with Layer 1 alone on the same reviewed attempts, the correlation rose from 0.19 to 0.73 and the error fell from 19.9 to 7.9 points.

**Answer to RQ2:** phoneme-level GOP scoring agrees substantially with teacher judgement for this population. It also reports which phoneme was weakest on each attempt, the information the adaptive mechanism uses.

## Reference recordings

The reference audit ran each of the 78 UK reference recordings through the GOP engine:

- **Before fixes:** 28 of 78 references decoded cleanly as their own word.
- **After phoneme-mapping fixes:** 46 of 78. The fixes were alternative tokens for consonant clusters, 0.3 s silence padding, and accent variants.
- **Later, stricter re-audit:** 32 of 78 (41%) still flagged as low quality.

The remaining failures come from the recordings themselves, which decode unrelated phonemes even though Whisper recognizes the correct word, so they need re-

recording. Because GOP scores the child's audio directly, poor references affect mainly the DTW part of the score (30%) and playback.

### Child-speech adaptation (RQ4)

**Data preparation (complete).** Nine classroom sessions (426.6 s of audio) were processed:

- **Segments:** 272 speech segments were found.
- **Speaker separation:** two pitch groups appeared in every recording. Adult cluster centres were 116–181 Hz and child cluster centres 296–340 Hz.
- **Child clips:** 167 segments (61%) were classed as child speech and cut into clips.
- **Target words:** found automatically for 50 clips. Of these, 35 came from the teacher's prompt, which proved more reliable than transcribing the child directly. The rest were left for the teacher to identify.
- **Production scoring of those 50 clips:** 45 were scored and 5 rejected as word mismatches. With room-tone padding, every clip passed the recording-quality check.

### Model training and evaluation

Table 6.7: *Adaptation model, held-out evaluation (grouped by session)*

| Model                            | Labelled clips | Error (points) | r    | Within 10 points | Pass/fail agreement |
|----------------------------------|----------------|----------------|------|------------------|---------------------|
| Baseline (current overall score) | 32             | 7.94           | 0.73 | 71.9%            | 87.5%               |
| Isotonic recalibration           | 32             | 11.46          | 0.65 | 71.9%            | 84.4%               |
| Ridge regression (all features)  | 32             | 13.43          | 0.57 | 62.5%            | 84.4%               |

## 6.2 Product and technical results

Table 6.8: *Status of each objective*

| Objective                               | Result                                                                                                                               |
|-----------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------|
| 1. Learning content                     | 78 items (26 letters, 52 words, 4 categories) with UK reference audio, pictures, phoneme mappings and difficulty profiles            |
| 2. Multimodal interface                 | Picture, reference audio for each phoneme, repeatable before assessment                                                              |
| 3. Speech acquisition and preprocessing | On-device recording (AAC, 16 kHz, mono) with a server-side quality check, and every in-app attempt in Dataset A passed through it    |
| 4. Pronunciation assessment             | Three-layer cascade producing word-level and phoneme-level scores ( $r = 0.73$ with teachers)                                        |
| 5. Reinforcement activities             | Listen-and-choose, Tap the Sounds (no speech needed, errorless), spoken word                                                         |
| 6. Adaptive learning                    | Next-word selection from score, weakest phoneme, difficulty and recurrence, with a targeted listening round when a weak sound recurs |
| 7. Teacher monitoring                   | Scores, weakest phoneme, replay, history, uncertainty-ranked review queue                                                            |
| 8. Evaluation                           | Agreement, Layer-1 and Layer-3 analyses completed (Section 6.1)                                                                      |
| 9. Child-speech adaptation              | Pipeline and model implemented, training pending labels                                                                              |

Table 6.9: *Technical results*

| Measure                                                                          | Result                          |
|----------------------------------------------------------------------------------|---------------------------------|
| Backend automated tests                                                          | 87 / 87 passed                  |
| Frontend pronunciation tests                                                     | 16 / 16 passed                  |
| Scoring time per attempt, model already loaded (development machine, all layers) | about 3.0 s (50 clips in 149 s) |
| Cold start, first attempt including model loading                                | about 12.8 s                    |
| Memory used by the GOP worker                                                    | about 1.3 GB                    |
| Attempts flagged for teacher review                                              | 26.1%                           |
| Agreement between backend and scikit-learn predictions                           | within 0.0001 points            |

Verification found and fixed five defects (Table 5.4): iOS replay, a double-tap recorder conflict, an ffmpeg overwrite hang, the GOP worker's Python environment, and missing-value handling in the model.

### 6.3 User or stakeholder validation

**Expert validation of scores.** Special education teachers reviewed 32 attempts through the review queue. Their scores are the reference standard behind the agreement results. On these attempts, system and teacher agreed on pass/fail 87.5% of the time and were within 10 points 71.9% of the time.

**Usability evaluation.** planned a usability evaluation with teachers. Insert the procedure and results here if one was carried out. Otherwise, report it as not completed and move it to future work.

### 6.4 Integration results

- **Integrated flow worked in practice.** All 245 attempts in Dataset A went through the complete integrated flow: teacher login with the platform's JWT, student selection from shared records, the pronunciation screens in the teacher navigation stack, scoring by the shared backend, and storage in the shared database. The flow therefore operated end to end throughout the evaluation period.
- **Scores written by the server.** All scores were written by the server, and saved results referred to their scored attempt by `result_id`.
- **Tablet build.** The iOS bundle for Expo Go was compiled without errors (11.4 MB in development mode) against the current code, and the backend health and authentication endpoints responded correctly over the local network.
- **Population field.** The student population field from the shared records drove the per-population analyses (Tables 6.1, 6.6) without any changes to the student data model.

## 6.5 Summary of findings

1. **RQ1: No.** Word-level MFCC-DTW against an adult UK reference, the approach originally proposed, did not reliably assess autistic children's pronunciation. It accepted no attempts (0/151), its correlation with teachers was not distinguishable from zero ( $r = 0.19$ ), and no accept threshold worked.
2. **RQ2: Yes.** Phoneme-level GOP scoring agreed substantially with teachers:  $r = 0.73$ , mean absolute error 7.9 points, 87.5% pass/fail agreement. It also identifies the weakest phoneme, which drives adaptation.
3. **RQ3: Yes, as a working system.** An adaptive pronunciation component with teacher review was built and used in 249 real attempts. It covers all eight original objectives. About a quarter of attempts were flagged for teacher judgement rather than presented as certain.
4. **Recalibration** from teacher reviews improved the internal adaptive score (error from 11.2 to 8.8), but not the score teachers see. ASD populations need more reviews before they get their own calibration curves.
5. **RQ4: pending.** The child-speech adaptation pipeline worked as designed: teacher and child separated cleanly by pitch, and 167 child clips were extracted. Whether the trained model improves agreement will be known after teacher labelling.
6. **Limits on these findings:** all results are preliminary. Only 32 attempts were reviewed, and they were chosen by an uncertainty-based review rather than at random. 25 of the 32 came from students with learning disabilities rather than ASD, and 41% of reference recordings need re-recording.

## 7 DISCUSSION

### 7.1 Interpretation of findings

**The phoneme-level scoring approach works for the target learners.** The component's scores agreed closely with teacher judgement:  $r = 0.73$ , a mean absolute error of 7.9 points on a 100-point scale, and 87.5% agreement on whether an attempt reached the 70-point threshold. In seven of eight attempts the system and the teacher reached the same decision. This shows that phoneme-level Goodness of Pronunciation scoring, built on a pretrained wav2vec2 recognizer and used without any additional training, can give teachers consistent, evidence-based pronunciation scores for children in special education settings.

The development process supports the design changes that were made. The proposal began with word-level MFCC-DTW similarity. Measuring it on real attempts showed why phoneme-level scoring was needed: DTW against an adult reference stayed in a narrow 60–80 band whatever the pronunciation quality ( $r = 0.19$  with teachers). This is consistent with the known acoustic differences between children's and adults' speech [17], [11]. Rather than discarding DTW, the component keeps it as a fast, accept-only first check. Because it can never fail a child on its own, no child was penalized for the voice difference. Moving from word-level to phoneme-level scoring raised agreement with teachers from  $r = 0.19$  to  $r = 0.73$  and cut the error from 19.9 to 7.9 points. This change came directly from evaluating the first design, which is how a design and development approach is meant to work.

**Teachers stay in control of uncertain cases.** The component flagged 26.1% of attempts for teacher review, about one in four. Confident results flow automatically, and uncertain ones reach the teacher rather than being presented as certain. This load is manageable in a classroom and keeps the teacher as the final judge, as the proposal intended.

The validation rules prevented scoring quality from getting worse. Layer 3 learns from teacher reviews, and its rules behaved as designed. They adopted a correction that reduced the error of the internal adaptive score from 11.2 to 8.8 points. They rejected

a correction to the displayed overall score, which was already well calibrated and would have become. In other words, a correction is applied only where the data supports it, and scores that are already accurate are left alone. Already working.

Ordinary classroom sessions can supply training data. The child-speech adaptation pipeline processed nine real classroom sessions. Teacher and child voices separated cleanly by pitch in every session (adult 116–181 Hz, child 296–340 Hz), as expected from the higher fundamental frequency of children's voices [17], producing 167 child utterances, and the teacher's own prompt proved a reliable way to identify each target word. Data for adapting the system to autistic children's speech can therefore be gathered from normal teaching, without a separate recording protocol.

## 7.2 Comparison with literature baselines and existing solutions

**Against existing applications.** Compared with the existing applications reviewed in Table 2.1:

- **Read Along** supports reading aloud, not the pronunciation of individual words.
- **ELSA Speak** gives phoneme feedback, but to adult learners, directly, with no teacher in the loop.
- **Proloquo2Go and Avaz** support communication, not pronunciation training.

None combines autism-sensitive activities, phoneme-level scoring, adaptation to recurring difficult sounds, and teacher review of uncertain results. The component does all four, although at the smaller scale and lower maturity of a research prototype.

## 7.3 Research contribution

1. **Evidence on scoring methods for this population.** The study shows, using real attempts, that phoneme-level GOP scoring agrees substantially with teacher judgement for children in special education, whereas word-level

template matching against an adult reference does not. This finding is relevant to anyone building pronunciation tools for autistic children.

2. **A child-safe scoring cascade.** A three-layer design in which a fast acoustic check may accept an attempt but never fail it, phoneme-level scoring handles the attempts it does not accept, and teacher review handles any that remain uncertain.
3. **Adaptation gated by evidence, with the teacher in the loop.** The scoring improves from teacher judgement, separately for each learner population, and applies a correction only when held-out validation shows it helps.
4. **A method for creating child speech data from classroom practice.** Prompt-and-repeat sessions are separated by voice pitch, aligned to target words through the teacher's prompt, and labelled blind. The same method could be reused to build pronunciation data for under-represented learners.

#### 7.4 Product contribution

- **A complete, working pronunciation module** covering all eight objectives, used for 249 real attempts by 16 students.
- **Child-centered interaction:**
  - No scores are shown to children.
  - Tap the Sounds, an errorless activity, works on days when the child does not speak.
  - A reduced-stimulation option.
  - Picture, audio and mouth-shape support before any assessment.
  - A recording-quality check asks the child to try again rather than scoring noise.
- **Teacher tools:**

- the weakest phoneme for each attempt
- recording replay
- session history
- a review queue that puts the most useful reviews first.
- **Privacy-preserving processing.** All speech recognition and scoring runs on the project's own server, so children's voices are never sent to third-party speech services. A retention script removes old audio while keeping the score records.
- Low running cost. No GPU and no paid speech API are needed. Scoring takes about 3 seconds per attempt on a standard laptop, which suits the budgets of special education schools.
- Testing: 87 backend tests and 16 frontend pronunciation tests, all passing.

## 7.5 Limitations and threats to validity

- **Evaluation sample.**
  - **Limitation:** 32 teacher-reviewed attempts from 16 students in one school. The reviews were chosen by uncertainty-based review, and most came from students with learning disabilities.
  - **Mitigation:** results are reported with confidence intervals and population breakdowns, and the review queue gives extra priority to under-represented populations, so ASD coverage grows as the system is used.
- **Independence of teacher reviews.**
  - **Limitation:** the in-app review screen shows the system's score, which may have influenced some teacher reviews, so the agreement figures should be read as an upper estimate.

- **Mitigation:** the classroom-recording labelling tool was designed blind (it hides the system's score), and the same approach is planned for the in-app review screen.
- **Reference recordings.**
  - **Limitation:** some adult UK reference recordings are of lower quality.
  - **Mitigation:** phoneme-mapping fixes raised the number of clean references from 28 to 46 of 78, and the remaining recordings have been identified for re-recording. GOP scores the child's own speech, so poorer references mainly affect the 30% DTW part of the score.
- **Reference standard.**
  - **Limitation:** pronunciation is judged against UK English references and the teacher's judgement, which may not fully reflect local intelligibility.
  - **Mitigation:** teacher review and per-population calibration let the scoring adapt to local judgement over time.
- **Adaptation model.**
  - **Limitation:** the child-speech adaptation model is implemented, but its training is pending teacher labelling of the classroom clips.
  - **Mitigation:** the model can only replace the current score once it has beaten it on held-out children, so it does not affect live scoring until then.

## 7.6 Implications

### For teachers and schools.

- The component gives teachers objective, phoneme-level evidence of each child's pronunciation and points directly to the sound that needs practice.
- Uncertain cases are left to the teacher.

- It runs on ordinary hardware without paid services, so special education schools in Sri Lanka could realistically adopt it.

**For designers of tools for autistic learners.**

- Phoneme-level scoring is a better basis than template matching for children's speech [17], [11].
- Detailed scores belong with the teacher, while children should get encouragement.
- An automated measure should not be allowed to fail a child on its own.
- A system that learns from teacher judgement, with safeguards, can improve without putting learners at risk.

**For research.**

- Pretrained phoneme recognizers already give useful agreement with teachers without any adaptation.
- Collecting data from ordinary classroom practice offers a practical way to build the child-speech corpora the field lacks, which could raise accuracy further for autistic learners.

## 8 CONCLUSION AND FUTURE DIRECTIONS

### 8.1 Conclusion

This research set out to help Grade 3 autistic children in Sri Lanka learn foundational English pronunciation through an adaptive, teacher-supported component. Following a design and development research approach, the Adaptive English Pronunciation Support Module was designed, built and evaluated as a React Native tablet application with a Node.js backend.

The component guides each child through autism-sensitive multimodal support (picture prompts, reference audio and phoneme-level mouth-shape guidance) and then through reinforcement activities, including an errorless sound-segmentation game that needs no speech. Spoken attempts are scored by a three-layer cascade:

- **Layer 1:** an accept-only MFCC-DTW comparison with UK reference audio.
- **Layer 2:** phoneme-level Goodness of Pronunciation scoring with a pretrained wav2vec2 model, supported by Whisper word verification.
- **Layer 3:** recalibration from teacher reviews, gated by validation.

An adaptive mechanism chooses each next word from the child's weakest phoneme, the word's difficulty and recurring difficulty patterns. Teachers receive phoneme-level results, recording replay, session history and a ranked review queue, while children receive only encouraging feedback.

Across 249 real attempts from 16 students, the component's scores agreed closely with teacher judgement:  $r = 0.73$ , a mean absolute error of 7.9 points, and 87.5% agreement on the pass/fail decision. The study also showed why phoneme-level scoring matters: moving from word-level template matching to GOP-based scoring raised agreement with teachers from  $r = 0.19$  to  $r = 0.73$ . The component therefore answers RQ1, RQ2 and RQ3. It also provides a working pipeline for adapting scoring to autistic children's speech from ordinary classroom recordings, for which RQ4 will be answered once teacher labelling is complete.

Overall, the research shows that a low-cost, locally hosted and teacher-supervised pronunciation component can give autistic learners structured, supportive practice, and give their teachers consistent, phoneme-level evidence for instruction.

## 8.2 Achievement of objectives

Table 8.1: *Achievement of objectives*

| # | Objective                            | Status                               | Evidence                                                                                                                                                       |
|---|--------------------------------------|--------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 1 | Collect and organize Grade 3 content | Achieved                             | 78 items (26 letters, 52 words in 4 categories) with UK reference audio, pictures, phoneme mappings and difficulty profiles                                    |
| 2 | Multimodal learning interface        | Achieved                             | Picture, reference audio and mouth-shape guidance for each phoneme, repeatable before assessment, and autism-friendly design with a reduced-stimulation option |
| 3 | Speech acquisition and preprocessing | Achieved                             | On-device AAC recording (16 kHz, mono) and a server-side quality check (duration, voiced speech, SNR, clipping)                                                |
| 4 | Pronunciation assessment             | Achieved and extended                | Word-level MFCC-DTW as proposed, extended with phoneme-level GOP and word verification ( $r = 0.73$ with teachers)                                             |
| 5 | Reinforcement activities             | Achieved                             | Listen-and-choose, Tap the Sounds (errorless, no speech needed) and spoken-word practice                                                                       |
| 6 | Adaptive learning mechanism          | Achieved                             | Next word chosen from score, weakest phoneme, difficulty and recurrence, with a targeted listening round when a weak sound recurs                              |
| 7 | Teacher-facing monitoring            | Achieved                             | Scores, weakest phoneme, recording replay, history and an uncertainty-ranked review queue                                                                      |
| 8 | Evaluation                           | Achieved                             | Agreement with teachers, Layer-1 and Layer-3 analyses, 87 backend and 16 frontend tests passing                                                                |
| 9 | Child-speech adaptation model        | Implemented, training pending labels | 167 child clips extracted from 9 classroom sessions, blind labelling tool, feature extraction, training and backend integration complete                       |

### 8.3 Individual research and product contributions

1. Evidence that phoneme-level GOP scoring agrees substantially with teacher judgement for children in special education, whereas word-level template matching against an adult reference does not.
2. A child-safe, three-layer scoring cascade in which a fast acoustic check may accept but never fail a child, and uncertain attempts escalate to phoneme-level scoring and then to the teacher.
3. Adaptation that learns from teacher judgement for each learner population, applying a correction only when held-out validation shows it improves accuracy.
4. A reusable method for building labelled child speech data from ordinary prompt-and-repeat classroom sessions, using pitch-based speaker separation, alignment through the teacher's prompt, and blind labelling.

#### Product contributions

1. A complete, working pronunciation module for autistic learners, used in 249 real attempts.
2. Child-centered interaction: encouragement instead of scores, errorless activities, and a reduced-stimulation option.
3. Teacher tools that turn scores into instructional evidence: weakest phoneme, replay, history and a review queue.
4. Speech processing that runs on the project's own server, protecting children's privacy, at low running cost with no GPU or paid speech services.

### 8.4 Future directions

1. **Train and deploy the child-speech adaptation model.** Complete teacher labelling of the 167 classroom clips, train and evaluate the model on held-out children, and activate it if it improves on the current scoring.

2. **Blind teacher reviews and random audit samples.** Start the in-app review score empty and hide the system's score, and route a small random share of confident attempts to review. Both give unbiased labels for calibration.
3. **More reviews from ASD learners.** Grow reviewed data for each ASD level until per-population calibration is supported (24 reviews each), then activate Layer 3 where validation passes.
4. **Better reference audio and child references.** Re-record the flagged UK references with lead-in silence, and add child reference recordings marked correct by teachers, so the DTW layer can compare children with children.
5. **Learn the scoring weights from data.** Replace the hand-set GOP–DTW blend and adaptive-score weights with values fitted to teacher labels once enough data exists.
6. **Adapt the phoneme model to children.** With a larger labelled corpus, fine-tune the wav2vec2 phoneme recognizer [8], [23] on children's and autistic children's speech.
7. **Usability study with teachers and learners.** Run a formal usability evaluation in classrooms, as planned in the proposal.
8. **Proposal features to extend.**
  - a. Lip-sync video guidance alongside the drawn mouth shapes.
  - b. A web-based teacher dashboard with category-wise performance charts.
  - c. A progress view for parents and guardians, who are stakeholders named in the proposal.
9. **Localization.** Add Sinhala (and Tamil) instructions translated by fluent speakers, while keeping English as the target language.
10. **Wider deployment.** Pilot the component in more special education schools and with other vocabulary sets, to test how well it generalizes beyond one school and the Grade 3 content.



## REFERENCES

- [1] American Psychiatric Association, *Diagnostic and Statistical Manual of Mental Disorders*, 5th ed., text rev. (DSM-5-TR). Washington, DC, USA: American Psychiatric Association Publishing, 2022.
- [2] H. Tager-Flusberg, R. Paul, and C. Lord, “Language and communication in autism,” in *Handbook of Autism and Pervasive Developmental Disorders*, vol. 1, 3rd ed., F. R. Volkmar, R. Paul, A. Klin, and D. Cohen, Eds. Hoboken, NJ, USA: Wiley, 2005, pp. 335–364.
- [3] H. Perera, K. Wijewardena, and R. Aluthwelage, “Screening of 18–24-month-old children for autism in a semi-urban community in Sri Lanka,” *Journal of Tropical Pediatrics*, vol. 55, no. 6, pp. 402–405, 2009.
- [4] A. W. Little, M. Shojo, U. Sonnadara, and H. Aturupane, “Teaching English as a second language in Sri Lankan primary schools: Opportunity and pedagogy,” *Language, Culture and Curriculum*, vol. 32, no. 2, pp. 113–127, 2019.
- [5] S. B. Davis and P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.
- [6] H. Sakoe and S. Chiba, “Dynamic programming algorithm optimization for spoken word recognition,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 26, no. 1, pp. 43–49, 1978.
- [7] S. M. Witt and S. J. Young, “Phone-level pronunciation scoring and assessment for interactive language learning,” *Speech Communication*, vol. 30, nos. 2–3, pp. 95–108, 2000.
- [8] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 12449–12460.
- [9] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, “Robust speech recognition via large-scale weak supervision,” in *Proc. 40th Int. Conf. Machine Learning (ICML)*, 2023, pp. 28492–28518.

- [10] Y. El Kheir, A. Ali, and S. A. Chowdhury, “Automatic pronunciation assessment – A review,” in *Findings of the Association for Computational Linguistics: EMNLP 2023*, Singapore, 2023, pp. 8304–8324.
- [11] M. Gerosa, D. Giuliani, S. Narayanan, and A. Potamianos, “A review of ASR technologies for children’s speech,” in *Proc. 2nd Workshop on Child, Computer and Interaction (WOCCI)*, 2009, pp. 1–8.
- [12] P. Rogerson-Revell, “Computer-assisted pronunciation training (CAPT): Current issues and future directions,” *RELC Journal*, vol. 52, no. 1, pp. 189–205, 2021.
- [13] O. Grynszpan, P. L. Weiss, F. Perez-Diaz, and E. Gal, “Innovative technology-based interventions for autism spectrum disorders: A meta-analysis,” *Autism*, vol. 18, no. 4, pp. 346–361, 2014.
- [14] J. Stephenson and M. Limbrick, “A review of the use of touch-screen mobile devices by people with developmental disabilities,” *Journal of Autism and Developmental Disorders*, vol. 45, no. 12, pp. 3777–3791, 2015.
- [15] Y. Getman, T. Grósz, M. Kurimo, and G. Salvi, “Non-native children’s automatic speech assessment challenge (NOCASA),” in *Proc. IEEE 35th Int. Workshop Machine Learning for Signal Processing (MLSP)*, 2025.
- [16] G. B. Mesibov, V. Shea, and E. Schopler, *The TEACCH Approach to Autism Spectrum Disorders*. New York, NY, USA: Springer, 2004.
- [17] S. Lee, A. Potamianos, and S. Narayanan, “Acoustics of children’s speech: Developmental changes of temporal and spectral parameters,” *Journal of the Acoustical Society of America*, vol. 105, no. 3, pp. 1455–1468, 1999.
- [18] D. D. Lewis and W. A. Gale, “A sequential algorithm for training text classifiers,” in *Proc. 17th Annu. Int. ACM SIGIR Conf. Research and Development in Information Retrieval*, 1994, pp. 3–12.
- [19] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York, NY, USA: Springer, 2009.
- [20] H. S. Terrace, “Discrimination learning with and without “errors”,” *Journal of the Experimental Analysis of Behavior*, vol. 6, no. 1, pp. 1–27, 1963.

- [21] S. M. Witt, “Use of speech recognition in computer-assisted language learning,” Ph.D. dissertation, Univ. Cambridge, Cambridge, U.K., 1999.
- [22] W. Hu, Y. Qian, and F. K. Soong, “An improved DNN-based approach to mispronunciation detection and diagnosis of L2 English speech,” in *Proc. Interspeech*, 2013, pp. 817–821.
- [23] Facebook AI, “wav2vec2-lv-60-espeak-cv-ft,” Hugging Face model card. [Online]. Available: <https://huggingface.co/facebook/wav2vec2-lv-60-espeak-cv-ft>
- [24] C. Saunders, A. Gammernan, and V. Vovk, “Ridge regression learning algorithm in dual variables,” in *Proc. 15th Int. Conf. Machine Learning (ICML)*, 1998, pp. 515–521.
- [25] B. Settles, “Active learning literature survey,” Computer Sciences Tech. Rep. 1648, Univ. Wisconsin–Madison, Madison, WI, USA, 2009.
- [26] J. Zhang et al., “speechocean762: An open-source non-native English speech corpus for pronunciation assessment,” in *Proc. Interspeech*, 2021, pp. 3710–3714.
- [27] L. van der Meer et al., “A systematic review of research on augmentative and alternative communication interventions for children with autism spectrum disorder,” *Augmentative and Alternative Communication*, vol. 28, no. 4, pp. 230–242, 2012.
- [28] G. Gerganov, “whisper.cpp: High-performance inference of OpenAI’s Whisper ASR model,” GitHub repository. [Online]. Available: <https://github.com/ggml-org/whisper.cpp>
- [29] M. Jones, J. Bradley, and N. Sakimura, “JSON Web Token (JWT),” IETF RFC 7519, May 2015.
- [30] *Systems and Software Engineering — Life Cycle Processes — Requirements Engineering*, ISO/IEC/IEEE 29148:2018, International Organization for Standardization, Geneva, Switzerland, 2018.
- [31] *Systems and Software Engineering — Systems and Software Quality Requirements and Evaluation (SQuaRE) — System and Software Quality Models*, ISO/IEC 25010:2011, International Organization for Standardization, Geneva, Switzerland, 2011.

- [32] Parliament of the Democratic Socialist Republic of Sri Lanka, *Personal Data Protection Act, No. 9 of 2022*, Colombo, Sri Lanka, 2022.
- [33] W3C, *Web Content Accessibility Guidelines (WCAG) 2.1*, W3C Recommendation, Jun. 5, 2018. [Online]. Available: <https://www.w3.org/TR/WCAG21/>

## GLOSSARY

*[Term - definition.]*

## **APPENDICES**

**Appendix A Research instruments or extended results**

**Appendix B Academic integrity and AI use disclosure**