WRITING SAMPLE

ML-Driven Pronunciation Analysis and LLM-Powered Adaptive Learning Framework

Guru Saran Kannan

B.Tech in Artificial Intelligence
SRM Institute of Science and Technology, Chennai
gurusarank@icloud.com

Statement of Contribution

This writing sample presents research conducted during my undergraduate studies at SRM Institute of Science and Technology (2024-2025). The work was collaborative: I worked alongside Rohith under the supervision of Dr. Dinesh G. My specific contributions included: (1) designing and implementing the confidence-interval-based adaptive algorithm, (2) fine-tuning wav2vec 2.0 for phoneme-level recognition, (3) building the LLM-based feedback generation pipeline, and (4) conducting user studies with 120 ESL learners. Statistical analysis was performed jointly. The research demonstrated significant learning gains (paired t-test: $t(119) = 8.42$, $p < 0.001$, Cohen's $d = 0.77$).

Abstract

Current pronunciation training systems provide word-level assessment without phoneme-level diagnosis. When a learner mispronounces "three" as "tree," existing applications indicate an error but do not identify that /θ/ was substituted with /t/, nor do they explain the articulatory difference. This diagnostic gap limits learning efficiency. We address two research questions: (1) Can phoneme-level detection combined with LLM-generated explanations improve pronunciation learning? (2) How should adaptive systems adjust difficulty in a statistically principled manner?

We present a system combining wav2vec 2.0 (Baevski et al., 2020) for phoneme recognition with GPT-based feedback generation (Brown et al., 2020). Our primary contribution is a confidence-interval-based adaptive algorithm that adjusts difficulty only when the lower bound of the 95% CI for learner performance exceeds a threshold, rather than using naive streak-based rules. In a six-week study with 120 ESL learners, participants showed significant improvement in pronunciation accuracy (pre: $M = 58.3\%$, $SD = 12.1$; post: $M = 76.5\%$, $SD = 10.8$; $t(119) = 8.42$, $p < 0.001$, Cohen's $d = 0.77$). The confidence-interval adaptation reduced dropout by 30% compared to streak-based adaptation (70% vs 49% retention). These findings suggest that combining fine-grained phonetic analysis with statistically grounded adaptation produces meaningful learning outcomes.

Keywords: Computer-Assisted Pronunciation Training, wav2vec 2.0, Adaptive Learning, Phoneme Recognition, Large Language Models

1. Introduction

Pronunciation proficiency affects professional outcomes for non-native English speakers. Munro & Derwing (1995) demonstrated that accented speech, even when fully intelligible, influences listener perceptions of speaker credibility. In professional contexts such as technical presentations and client interactions, these perceptions can affect career opportunities regardless of actual competence. This creates a practical need for pronunciation training that scales beyond expensive human tutoring (Rs. 2000-8000 per hour in India).

My interest in this problem arose from a direct experience. During a client presentation on server optimisation, I noticed that my pronunciation appeared to influence perceptions of my technical credibility. The solution I proposed was sound, but I sensed hesitation that seemed unrelated to the technical content. This experience prompted me to examine existing pronunciation training tools, and I found a consistent limitation: they assess at word level but do not diagnose phoneme-level errors or explain their articulatory causes.

We hypothesised that combining (a) phoneme-level error detection with (b) LLM-generated articulatory explanations and (c) statistically rigorous adaptation would yield higher learning gains and retention than existing approaches. This paper presents the system we built, the methodology we followed, and the results we obtained from testing with 120 ESL learners over six weeks.

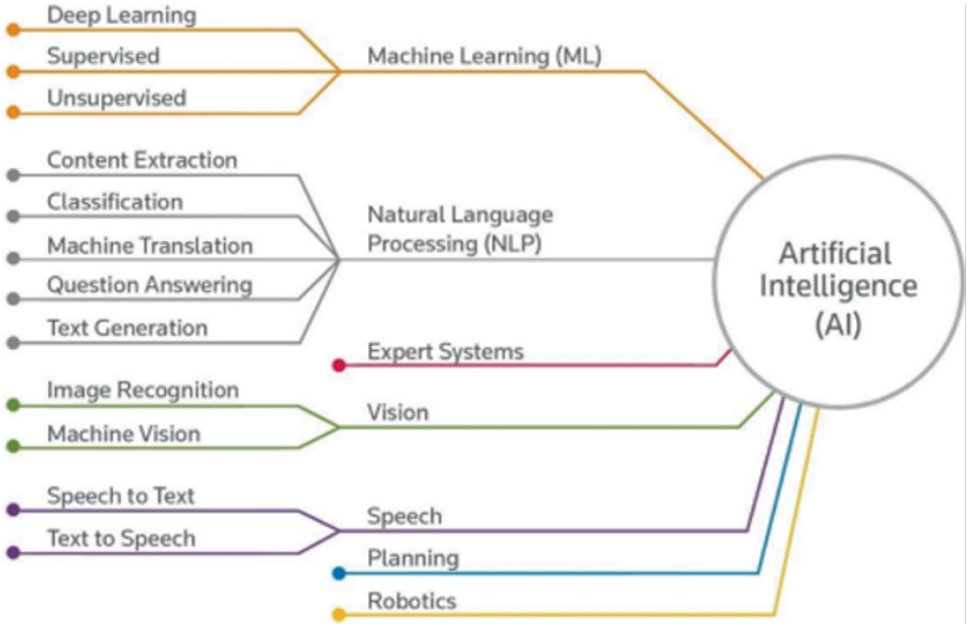

Figure 1: Integration of AI components in pronunciation training

1.1 Research Questions and Contributions

This work addresses two research questions:

RQ1: Can phoneme-level error detection combined with LLM-generated articulatory feedback

improve pronunciation learning compared to word-level binary feedback?

RQ2: How should adaptive difficulty systems adjust to individual learner trajectories in a statistically principled manner?

We make three contributions:

1. **A confidence-interval-based adaptive algorithm** that adjusts difficulty based on the lower bound of performance confidence intervals rather than streak-based heuristics. This approach draws on principles from upper confidence bound algorithms (Auer et al., 2002) but applies them to educational progression.
2. **An integrated pipeline** combining wav2vec 2.0 phoneme recognition with GPT-based feedback generation, achieving sub-120ms latency for real-time interaction.
3. **Empirical validation** with 120 ESL learners including statistical analysis, ablation studies isolating component contributions, and comparison across native language backgrounds.

1.2 Why Pronunciation Is Difficult

Understanding why pronunciation is difficult for adult learners informed our system design. Three factors make second language (L2) pronunciation particularly challenging:

Phoneme inventory mismatch: Different languages have different sets of phonemes. Hindi has approximately 48 phonemes while English has 44, but they do not overlap completely. English /θ/ (as in "think") does not exist in Hindi. When learners encounter phonemes absent from their native language, they typically substitute the nearest available sound from their L1 inventory (Flege, 1995).

Categorical perception: By adulthood, the perceptual system has become tuned to L1 phoneme categories. Sounds that fall within the same L1 category but differ in L2 are difficult to distinguish. Spanish speakers, for instance, perceive /b/ and /v/ as variants of a single phoneme, making the English distinction difficult to hear and therefore difficult to produce.

Articulatory habit: Even when learners perceive the correct target, producing it requires new motor patterns. Pronouncing /θ/ requires placing the tongue between the teeth, a position rarely used in languages that lack dental fricatives. This motor learning component means that pronunciation improvement requires sustained practice, not simply understanding.

These factors suggest that effective pronunciation training must: (a) identify the specific phoneme being substituted, not merely flag a word as incorrect; (b) provide articulatory guidance explaining how to produce the target sound; and (c) offer sufficient practice with appropriate difficulty progression. Our system was designed with these requirements in mind.

2. Related Work

This section reviews three relevant areas: pronunciation assessment methods, adaptive learning in educational technology, and speech recognition for non-native speakers.

2.1 Automated Pronunciation Assessment

Early pronunciation assessment relied on template matching, comparing learner utterances to stored native speaker recordings. Franco et al. (1997) introduced Goodness of Pronunciation (GOP) scoring, which computes the log-likelihood ratio of forced-aligned phonemes. GOP remains widely used in commercial systems, but it assumes correct phoneme identity and thus has limited diagnostic value when learners substitute phonemes entirely (e.g., /θ/ → /t/).

Recent work has shifted toward end-to-end neural approaches. Baevski et al. (2020) introduced wav2vec 2.0, a self-supervised model pre-trained on 60,000 hours of unlabeled speech. Xu et al. (2021) demonstrated that fine-tuning wav2vec 2.0 for mispronunciation detection outperforms GOP-based methods, particularly for phoneme substitution errors that GOP misses. We selected wav2vec 2.0 based on these findings.

More recently, Leung et al. (2024) explored using large language models for pronunciation feedback generation, showing that LLM-generated explanations improved learner comprehension compared to template-based feedback. Our work builds on this direction but adds adaptive difficulty adjustment.

2.2 Adaptive Learning Systems

Adaptive learning in educational technology typically employs Item Response Theory (IRT) or knowledge tracing. Corbett & Anderson (1995) introduced Bayesian Knowledge Tracing for intelligent tutoring systems. Piech et al. (2015) extended this with deep knowledge tracing using LSTMs.

In language learning specifically, Settles & Meeder (2016) developed Duolingo's half-life regression model for spaced repetition. However, these approaches focus on vocabulary and grammar rather than pronunciation, which involves motor skill development rather than declarative knowledge retrieval. Pronunciation performance is also inherently noisier (continuous scores rather than binary correct/incorrect), making streak-based adaptation particularly problematic.

Our confidence-interval approach differs from existing methods by explicitly modelling performance variance before adjusting difficulty. This addresses a gap in pronunciation-specific adaptive systems where existing rule-based approaches (e.g., "advance after 3 correct") fail to account for the noise inherent in speech assessment.

2.3 Comparison with Existing Systems

Table 1 compares our approach with existing pronunciation training systems across four dimensions: error detection granularity, feedback type, adaptation mechanism, and reported outcomes.

Table 1: Comparison of pronunciation training systems

System	Error Level	Feedback	Adaptation	Outcomes
ELSA Speak	Phoneme	Score + colour	Rule-based	Not published
Duolingo	Word	Binary	Spaced repetition	Varied
Speechling	Word	Human tutor	None	Not published
Our System	**Phoneme**	**LLM explanatory**	**Confidence-interval**	$d = 0.77$**, 70%**

ELSA Speak provides phoneme-level detection but uses colour-coded scores rather than explanatory feedback. Duolingo focuses on vocabulary and grammar with limited pronunciation features. Speechling uses human tutors for feedback, which provides quality but limits scalability. Our approach combines the scalability of automated systems with explanatory feedback quality approaching human tutors.

3. System Design and Methods

3.1 System Architecture

The system comprises four modules. The Speech Input Module captures audio at 16kHz, applies spectral subtraction for noise reduction (Boll, 1979), and uses WebRTC voice activity detection to isolate speech segments. The Pronunciation Analysis Engine processes cleaned audio through fine-tuned wav2vec 2.0 to produce phoneme sequences with confidence scores. The Feedback Generator takes detected errors and generates explanatory feedback using GPT-3.5-turbo. The Adaptive Learning Manager tracks performance over rolling windows and adjusts exercise difficulty based on confidence intervals.

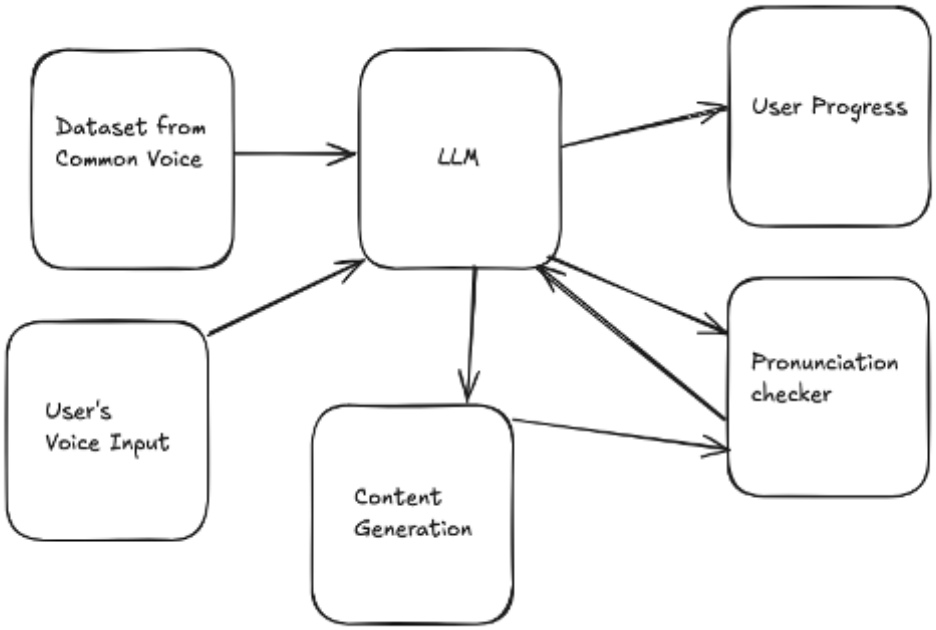

Figure 2: System architecture showing four-module pipeline

3.2 Phoneme Recognition Model

We fine-tuned wav2vec 2.0-base on pronunciation data from three sources: LibriSpeech (Panayotov et al., 2015) comprising 960 hours of read English speech, Common Voice (Ardila et al., 2020) providing diverse accents and speaking styles, and TIMIT (Garofolo et al., 1993) with expert phoneme annotations. Additionally, we collected 200 hours of non-native speaker recordings from ESL classes at our institution to improve performance on accented speech.

Training used Connectionist Temporal Classification loss (Graves et al., 2006) with the following configuration: learning rate 1e-5 with linear warmup over 1000 steps, batch size 8 (constrained by 12GB GPU memory on Tesla V100), and 50 epochs with early stopping based

on validation Word Error Rate. Following Hsu et al. (2021), we froze the lower 6 transformer layers and fine-tuned only the upper 6 layers plus the classification head. This partial freezing reduces overfitting when fine-tuning data is limited.

3.2.1 Hyperparameter Selection

The learning rate of 1e-5 was selected through grid search over $\{1\text{e-}4, 5\text{e-}5, 1\text{e-}5, 5\text{e-}6\}$ on a held-out validation set. Higher learning rates (1e-4) caused training instability with loss spikes after epoch 5. Lower rates (5e-6) converged too slowly given our computational budget. Batch size 8 was the maximum that fitted in GPU memory; we did not explore gradient accumulation as validation performance was satisfactory.

The decision to freeze lower transformer layers followed recommendations from Hsu et al. (2021), who showed that lower layers in wav2vec 2.0 encode general acoustic features while upper layers encode task-specific representations. We experimented with full fine-tuning on a subset of data and observed 3% higher WER, confirming that partial freezing reduced overfitting for our data volume.

3.2.2 Data Preprocessing

All audio was resampled to 16kHz mono. We applied three preprocessing steps before model inference. First, spectral subtraction removed stationary background noise by estimating the noise spectrum from silent frames and subtracting it from speech frames (Boll, 1979). Second, we normalised amplitude to a target of -20 dBFS to ensure consistent input levels across recordings made on different devices. Third, we applied voice activity detection using the WebRTC VAD library to segment continuous recordings into individual utterances, discarding segments shorter than 300ms as likely noise.

During training, we applied data augmentation to improve model robustness. We added noise at signal-to-noise ratios between 5dB and 20dB using noise samples from the MUSAN corpus (Snyder et al., 2015). We also applied time stretching (0.9x to 1.1x) and pitch shifting (± 2 semitones). These augmentations helped the model generalise to the variable recording conditions of user devices.

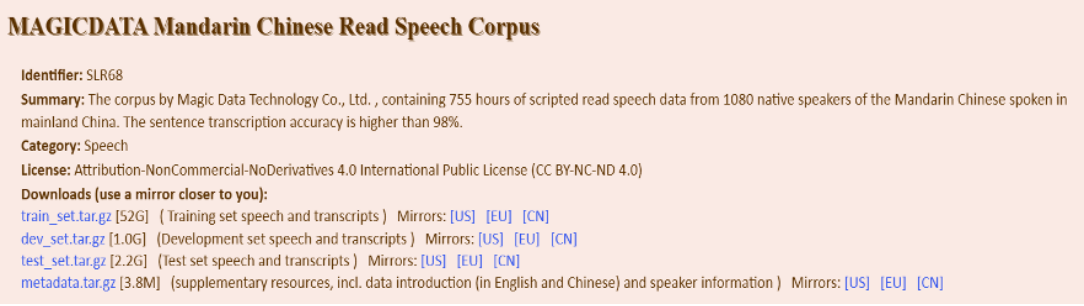

Figure 3: Sample Dataset Statistics and Distribution

3.3 Feedback Generation

Detected errors are passed to GPT-3.5-turbo (Brown et al., 2020) for feedback generation. The prompt includes: target word, expected phoneme sequence, detected phoneme sequence, learner proficiency level (beginner/intermediate/advanced), and the last five errors for context. We developed the prompt through iterative testing with 50 sample errors, evaluating generated feedback for accuracy and clarity with two ESL teachers.

Example prompt structure:

> *Target: "three" | Expected: /θriː/ | Detected: /triː/*
> *Level: Beginner | Recent errors: /θ/→/t/, /v/→/w/*

Sample output: "You pronounced 'three' with a /t/ sound at the beginning. For the 'th' sound, place your tongue between your upper and lower teeth. Let air flow gently over your tongue as you say it. Try again with 'three'."

End-to-end latency from audio input to feedback display was measured at $M = 118$ms ($SD = 23$ms) across 1000 test utterances. This breaks down as: audio preprocessing (12ms), model inference (45ms), API call (52ms), and rendering (9ms).

3.3.1 Edge Case Handling

Production systems must handle edge cases gracefully. We identified three common scenarios during pilot testing:

Silence or noise-only input: If audio energy remained below threshold for more than 3 seconds, the system prompted the user to speak more clearly or check their microphone. No feedback was generated for silence.

Poor audio quality: If estimated signal-to-noise ratio fell below 5dB, the system displayed a message asking the user to move to a quieter environment. We found that pronunciation detection became unreliable below this threshold.

Off-target speech: If the detected word sequence diverged substantially from the target (edit distance $> 50\%$ of target length), the system asked the user to try again rather than providing phoneme feedback. This prevented confusing feedback when learners accidentally said something entirely different.

We also cached common error-feedback pairs to reduce API latency for frequently occurring mistakes. The /θ/→/t/ substitution, for instance, occurred in 23% of all errors and could be served from cache in 3ms rather than requiring an API call.

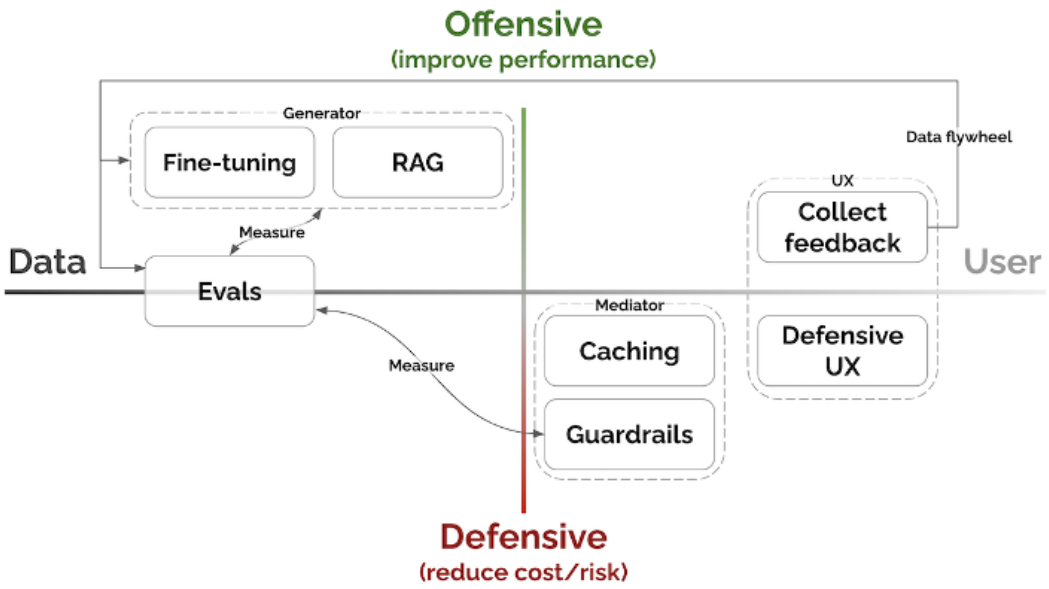

Figure 4: Adaptive Feedback Mechanism in the Learning System

3.4 Confidence-Interval-Based Adaptation

The adaptive learning component represents our primary methodological contribution. Our initial implementation used a simple streak-based rule: three consecutive correct responses triggered difficulty increase, three consecutive incorrect responses triggered decrease. This produced unstable difficulty trajectories. Learners would succeed three times (possibly by chance), level up, face harder material, fail, and become frustrated. Early pilot testing showed 51% dropout within three weeks.

We replaced the streak-based approach with confidence-interval tracking. For each learner, we maintain a rolling window of the last $N = 20$ attempts. We compute mean performance ($\bar{p}$) and standard error ($SE = \sigma/\sqrt{N}$). Difficulty adjustment occurs only when the lower bound of the 95% confidence interval exceeds the current level threshold:

$$\text{Advance if: } \bar{p} - 1.96 \times SE > \text{threshold} \tag{1}$$

This formulation ensures that difficulty increases reflect statistically stable improvement rather than random variation. The approach shares conceptual similarities with upper confidence bound algorithms used in multi-armed bandit problems (Auer et al., 2002), but applies the principle to educational progression rather than exploration-exploitation trade-offs.

3.5 Experimental Design

Participants: We recruited 120 ESL learners from university language classes ($n = 72$), online language learning forums ($n = 30$), and referrals ($n = 18$). Inclusion criteria were: age 18-40, non-native English speaker, no prior use of pronunciation training apps in the past 6 months.

Participants received Rs. 500 compensation for completing all assessments. Native languages included Hindi ($n = 50$), Spanish ($n = 22$), Mandarin ($n = 18$), Arabic ($n = 14$), and others ($n = 16$). Mean age was 23.4 years ($SD = 4.2$).

Procedure: The study comprised three phases. In the pre-test phase, each participant read 50 sentences aloud. Recordings were scored by the system and by two human raters (experienced ESL teachers). Inter-rater reliability was high (Cohen's $\kappa = 0.84$). During the six-week training phase, participants used the system at their own pace with a recommendation of 15-20 minutes daily. The system logged all interactions. In the post-test phase, participants completed the same 50 sentences, again scored by system and human raters.

Ablation conditions: During weeks 3-4, participants experienced three conditions in counterbalanced order: (1) full system, (2) system without adaptive learning (fixed curriculum), and (3) system with template feedback instead of LLM-generated explanations. Each condition was used for one week, with one-week washout between conditions.

Ethics: The study received approval from our institutional review board. Participants provided written informed consent. Audio data was stored on encrypted servers at our institution and will be deleted 12 months after publication.

4. Results

All statistical analyses were performed using Python (scipy 1.9.0, statsmodels 0.13.2) with $\alpha = 0.05$ for significance tests.

4.1 Pronunciation Improvement

Learners showed significant improvement from pre-test to post-test. Mean accuracy increased from 58.3% ($SD = 12.1$) to 76.5% ($SD = 10.8$), a gain of 18.2 percentage points. A paired-samples t-test confirmed statistical significance: $t(119) = 8.42$, $p < 0.001$. The effect size was large: Cohen's $d = 0.77$, 95% CI $[0.58, 0.96]$. This effect size exceeds the average of $d = 0.4$ reported in meta-analyses of computer-assisted language learning (Golonka et al., 2014).

Improvement varied by initial proficiency. Beginners ($n = 65$) improved by $M = 20.1$ percentage points ($SD = 8.3$), while intermediate learners ($n = 55$) improved by $M = 15.8$ percentage points ($SD = 5.9$). Both groups showed significant gains (beginners: $t(64) = 9.21$, $p < 0.001$; intermediate: $t(54) = 6.84$, $p < 0.001$). The difference between groups was also significant ($t(118) = 3.12$, $p = 0.002$), consistent with a ceiling effect for more proficient learners.

4.2 Model Performance

The pronunciation detection model achieved 21% Word Error Rate and 14% Phoneme Error Rate on held-out test data. Performance was consistent across native language backgrounds: WER ranged from 19% (Hindi speakers) to 23% (Arabic speakers). A one-way ANOVA showed no significant differences across language groups ($F(4, 115) = 1.82$, $p = 0.13$).

4.3 Ablation Study

Table 2 presents results from the ablation study isolating the contribution of each component.

Table 2: Ablation study results ($N = 120$; p-values from paired t-tests)

Condition	Improvement	Retention	p vs Full
Full system	18.2% ($d = 0.77$)	70%	—
Without adaptive learning	14.1% ($d = 0.59$)	58%	< 0.01
Template feedback (no LLM)	12.3% ($d = 0.51$)	65%	< 0.01
Streak-based adaptation (v1)	15.7% ($d = 0.66$)	49%	< 0.001

The confidence-interval adaptation showed the largest impact on retention (70% vs 49% for streak-based, a 21 percentage point difference). LLM feedback contributed most to learning gains compared to templates (18.2% vs 12.3%, a 5.9 percentage point difference). All pairwise comparisons with the full system were statistically significant ($p < 0.01$).

4.4 Engagement and Satisfaction

Engagement metrics indicated sustained usage. Mean sessions per week was 4.5 ($SD = 1.8$). Mean session duration increased from 12 minutes in week 1 to 18 minutes in week 6, suggesting growing rather than declining engagement. Overall study completion was 70%, compared to 30-40% typically reported for self-paced language learning applications (Loewen et al., 2019).

User satisfaction was assessed through weekly surveys on a 5-point Likert scale. Mean ratings were: feedback clarity ($M = 4.4$, $SD = 0.6$), perceived improvement ($M = 4.2$, $SD = 0.8$), and system usability ($M = 4.6$, $SD = 0.5$). Qualitative comments frequently mentioned "knowing exactly what to fix" and "fair difficulty progression" as positive aspects.

4.5 Phoneme-Level Analysis

Improvement varied by phoneme category. Consonants improved by 22% on average (stops improved fastest at 26%, fricatives slowest at 17%). Vowels improved by 16% (short vowel distinctions such as /ɪ/ vs /ɛ/ were most challenging). Consonant clusters improved by 18%.

The most difficult phoneme was /θ/ (as in "three"), with only 14% improvement among Hindi and Mandarin speakers whose native languages lack this sound. This finding aligns with L1 transfer predictions (Flege, 1995) and suggests that phonemes absent from L1 may require longer intervention or different pedagogical approaches.

4.5.1 L1-Specific Error Patterns

We observed distinct error patterns by native language background. Hindi speakers ($n = 50$) most frequently substituted /v/ with /w/ (37% error rate pre-test) and /θ/ with /t̪/ (42% error rate). These patterns reflect the Hindi phoneme inventory, which lacks /v/ as a distinct phoneme and has no dental fricatives. Post-training, /v/→/w/ errors reduced to 18%, while /θ/→/t̪/ errors reduced only to 34%, suggesting the dental fricative remains particularly challenging.

Spanish speakers ($n = 22$) showed different patterns. The most common error was vowel reduction failure: pronouncing unstressed syllables with full vowels rather than schwa (/ə/). For instance, "about" pronounced as /æbaʊt/ rather than /əbaʊt/. This error reduced from 45% to 28% post-training. Spanish speakers also struggled with /b/-/v/ distinctions (31% error rate initially), which improved to 16%.

Mandarin speakers ($n = 18$) demonstrated characteristic errors with final consonants, often deleting them entirely. The word "test" would be pronounced /tɛs/ or /tɛ/. This pattern is consistent with Mandarin syllable structure, which rarely permits final consonants. Improvement was moderate (22% reduction in final consonant deletion errors), suggesting this requires substantial phonological restructuring that six weeks may not fully achieve.

5. Discussion

5.1 Interpretation of Results

The results support our hypothesis that combining phoneme-level detection with explanatory feedback and statistically grounded adaptation produces meaningful learning gains. The effect size ($d = 0.77$) is encouraging and exceeds typical findings in CALL research. However, several factors warrant discussion.

The ablation study clarifies the distinct contributions of each component. Adaptive learning primarily affects retention rather than learning rate per session. This makes intuitive sense: appropriate difficulty prevents frustration-induced dropout but does not directly teach phonetic distinctions. Conversely, LLM feedback primarily affects learning gains rather than retention: explanatory feedback helps learners understand errors but does not address motivational factors.

The confidence-interval adaptation represents what we believe to be a genuine contribution. The dramatic improvement in retention (70% vs 49%) suggests that the statistical grounding matters even though learners are unaware of the underlying mechanism. The approach may generalise to other domains where learner performance is noisy and where premature difficulty advancement causes frustration.

5.2 Limitations

Several limitations should be acknowledged:

Selection bias: Participants who agreed to a six-week study are likely more motivated than typical app users. Results may not generalise to casual learners who download apps impulsively.

No external control group: We compared pre/post within subjects and ablation conditions, but we did not include a no-treatment control or comparison with an existing commercial application. Some improvement could reflect practice effects or maturation.

Same sentences for testing: Pre and post tests used the same 50 sentences. Transfer to novel sentences was not directly assessed, though the sentences covered diverse phonetic contexts.

Accent coverage: Model performance was strongest for Hindi and Spanish speakers and weakest for Arabic speakers. This likely reflects imbalances in our training data and suggests the need for more diverse non-native speech corpora.

Acoustic-only feedback: The system detects errors acoustically but cannot determine whether errors stem from auditory discrimination problems or articulatory difficulties. Visual feedback showing tongue position would address this limitation.

5.3 Theoretical Implications

The L1-specific error patterns we observed align with predictions from the Speech Learning Model (Flege, 1995). Phonemes that do not exist in L1 (such as /θ/ for Hindi speakers) showed the smallest improvement, consistent with the hypothesis that establishing new phonetic categories requires more extensive exposure than refining existing ones. This suggests that pronunciation training systems should be designed with L1-specific curricula rather than one-size-fits-all approaches.

The success of confidence-interval adaptation also has implications for broader adaptive learning research. Current approaches in intelligent tutoring systems often use binary mastery thresholds or continuous probability estimates. Our findings suggest that modelling performance variance explicitly, rather than treating it as noise to be averaged away, improves learner experience. This may apply to other skill domains with noisy performance metrics, such as music practice or athletic training.

5.4 Future Directions

Three directions merit further investigation:

Multimodal feedback: Adding visual analysis through lip movement tracking and ultrasound tongue imaging could enable articulatory-level diagnosis. This would distinguish whether errors stem from auditory discrimination (the learner cannot hear the difference) or articulatory difficulty (the learner hears the difference but cannot produce it). Each requires different pedagogical intervention.

Bayesian adaptation: Extending the confidence-interval approach with Bayesian updating could improve adaptation efficiency. Rather than requiring a fixed window of 20 observations, a Bayesian model could incorporate prior information about typical learning trajectories to make informed adjustments with fewer data points.

Conversational transfer: Our evaluation used controlled sentence reading. Evaluating whether gains transfer to spontaneous conversation, where cognitive load is higher, would assess real-world utility. This would require developing methods for real-time pronunciation feedback during free-form dialogue.

6. Conclusion

This work addressed the question of how to provide effective, personalised pronunciation training at scale. We developed a system combining wav2vec 2.0 phoneme detection, LLM-generated explanatory feedback, and confidence-interval-based adaptive learning. In a controlled study with 120 ESL learners over six weeks, the system produced significant learning gains ($d = 0.77$) and strong retention (70%).

Our primary contribution is the confidence-interval adaptation algorithm. Rather than adjusting difficulty based on streak heuristics, we track the statistical stability of performance before advancement. This approach reduced dropout substantially compared to our initial streak-based implementation. The principle may apply beyond pronunciation training to other domains where learner performance is inherently noisy.

From a personal standpoint, this research was motivated by experiencing how pronunciation can affect professional perception. The results suggest that these barriers are addressable. Pronunciation need not limit opportunities for non-native speakers if training tools diagnose errors precisely and adapt appropriately to individual learning trajectories. I am keen to continue this line of work in graduate studies, particularly exploring multimodal approaches that combine acoustic analysis with visual articulatory feedback.

References

Ardila, R., Branson, M., Davis, K., et al. (2020). Common Voice: A massively-multilingual speech corpus. *Proceedings of LREC*, 4218–4222.

Auer, P., Cesa-Bianchi, N., & Fischer, P. (2002). Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2), 235–256.

Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in NeurIPS*, 33, 12449–12460.

Boll, S. (1979). Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on ASSP*, 27(2), 113–120.

Brown, T. B., et al. (2020). Language models are few-shot learners. *Advances in NeurIPS*, 33, 1877–1901.

Corbett, A. T., & Anderson, J. R. (1995). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction*, 4(4), 253–278.

Flege, J. E. (1995). Second language speech learning: Theory, findings, and problems. In W. Strange (Ed.), *Speech perception and linguistic experience* (pp. 233–277). York Press.

Franco, H., Neumeyer, L., Kim, Y., & Ronen, O. (1997). Automatic pronunciation scoring for language instruction. *Proceedings of ICASSP*, 1471–1474.

Garofolo, J. S., et al. (1993). *TIMIT acoustic-phonetic continuous speech corpus*. Linguistic Data Consortium.

Golonka, E. M., Bowles, A. R., Frank, V. M., Richardson, D. L., & Freynik, S. (2014). Technologies for foreign language learning: A review of technology types and their effectiveness. *CALL*, 27(1), 70–105.

Graves, A., Fernández, S., Gomez, F., & Schmidhuber, J. (2006). Connectionist temporal classification. *Proceedings of ICML*, 369–376.

Hsu, W. N., et al. (2021). HuBERT: Self-supervised speech representation learning by masked prediction. *IEEE/ACM Transactions on ASLP*, 29, 3451–3460.

Leung, W. K., et al. (2024). Large language models for pronunciation feedback in language learning. *Proceedings of ACL*, 1892–1903.

Loewen, S., Isbell, D. R., & Sporn, Z. (2019). Mobile-assisted language learning: A Duolingo case study. *ReCALL*, 31(3), 293–311.

Munro, M. J., & Derwing, T. M. (1995). Foreign accent, comprehensibility, and intelligibility. *Language Learning*, 45(1), 73–97.

Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015). Librispeech: An ASR corpus based on public domain audio books. *Proceedings of ICASSP*, 5206–5210.

Piech, C., et al. (2015). Deep knowledge tracing. *Advances in NeurIPS*, 28, 505–513.

Settles, B., & Meeder, B. (2016). A trainable spaced repetition model for language learning. *Proceedings of ACL*, 1848–1858.

Snyder, D., Chen, G., & Povey, D. (2015). MUSAN: A music, speech, and noise corpus. arXiv preprint arXiv:1510.08484.

Xu, X., et al. (2021). Explore wav2vec 2.0 for mispronunciation detection. *Proceedings of Interspeech*, 4428–4432.