INTERNATIONAL PREMIER JOURNAL OF LANGUAGES & LITERATURE (IPJLL)
VOLUME 4, ISSUE 1, 2026 p-ISSN: 3007-2336 e-ISSN: 3007-2344

ASR-BASED FEEDBACK AND PRONUNCIATION ACCURACY: A QUASI-EXPERIMENTAL STUDY IN PAKISTAN

Syeda Noor Fatima syedanoor7521@gmail.com	BS Graduate, Department of English, University of Education, Lahore, Pakistan.
Sadia Ghaffar sadiaghafar@gmail.com	BS Graduate, Department of English, University of Education, Lahore, Pakistan.
Sidra Bukhari sidra.bukhari@ue.edu.pk	Assistant Professor, Department of English, University of Education, Lahore, Pakistan.
Dr. Tayyaba Yasmin chairenglish@ue.edu.pk	Associate Professor, Department of English, University of Education, Lahore, Pakistan.

Abstract

This study investigates the effectiveness of Automatic Speech Recognition (ASR) technology in improving the pronunciation accuracy of English as a Second Language (ESL) learners at a public-sector university in Lahore, Pakistan. Motivated by persistent challenges related to the absence of individualized feedback in large-enrollment classrooms and the documented gap in research addressing Pakistani EFL contexts, the study employs a quasi-experimental mixed-methods design with 40 participants (N = 40) randomly assigned to an experimental group (ASR-based instruction) or a control group (conventional teacher-mediated instruction). Quantitative data were collected at three time points: pre-test, immediate post-test, and delayed post-test, while qualitative data were obtained through semi-structured interviews. Independent samples t-tests, one-way ANOVA, and effect size analysis (Cohen's d) revealed a statistically significant and large-effect advantage for the experimental group (mean gain = 24.25 points; $d = 1.45$; $p < .001$). The delayed post-test confirmed superior long-term phonological retention in the experimental group (mean decay = −2.25 vs. −4.15 for the control group). Qualitative analysis yielded three emergent themes: the enabling role of immediate corrective feedback, a marked reduction in speaking anxiety attributable to the non-evaluative nature of ASR, and minor technical concerns about acoustic interference. Grounded in Schmidt's (1990) Noticing Hypothesis and Krashen's (1982) Affective Filter Hypothesis, these findings suggest that explicit ASR-based feedback can effectively bridge the input–output gap for ESL learners in the Pakistani higher education context.

Keywords: *Affective Filter, Automatic Speech Recognition (ASR), Capt, Corrective Feedback, ESL, Pakistan, Pronunciation Accuracy, Quasi-Experimental*

Corresponding Author: Syeda Noor Fatima (BS Graduate, Department of English, University of Education, Lahore, Pakistan)
Email: syedanoor7521@gmail.com

1. Introduction

English language learning in Pakistan presents a distinctive set of pedagogical challenges. In higher education settings, large class sizes, often exceeding 40 students per section, severely constrain teachers' capacity to provide the individualized, timely pronunciation feedback necessary for phonological development. Conventional classroom instruction frequently relies on implicit corrective feedback, which research has shown to be insufficient for learners who fail to consciously register the discrepancy between their production and target norms (Schmidt, 1990). Without such noticing, phonological errors fossilize, becoming resistant to later remediation.

Computer-Assisted Pronunciation Training (CAPT) powered by ASR technology has emerged internationally as a promising solution to this challenge. Unlike teacher-mediated feedback, which is inherently delayed, inconsistent, and constrained by classroom dynamics, ASR systems provide immediate, explicit, and repeatable phonological feedback. Learners can practice targeted sounds privately, iterating without the social anxiety associated with public correction (Lee & Lee, 2023). Comparative studies have consistently reported that learners who receive ASR-based feedback outperform those in conventional instruction, particularly at the intermediate proficiency level (Ngo et al., 2024).

Despite a growing global literature on CAPT, the Pakistani ESL context remains understudied. Existing research is concentrated in East Asian and European settings and does not account for the distinct sociolinguistic features of Pakistani English learners, including phonological interference from regional languages such as Urdu, Punjabi, Sindhi, and Pashto, as well as context-specific affective and infrastructural constraints. This study addresses that gap by examining the impact of ASR-based feedback on pronunciation accuracy among university-level ESL learners in Lahore, Pakistan, using a rigorous quasi-experimental mixed-methods design validated through Shadish et al.'s (2002) standards for causal inference.

1.1. Research Questions

This study addresses the following research questions:

1. What is the impact of ASR technology on pronunciation accuracy among Pakistani university students relative to conventional feedback methods?
2. How do students perceive the role of ASR technology in alleviating speaking anxiety?

2. Literature Review

2.1. Theoretical Framework: The Noticing Hypothesis

This study is theoretically grounded in Schmidt's (1990) Noticing Hypothesis, one of the most widely cited theoretical constructs in second language acquisition (SLA).

Schmidt argued that linguistic input cannot become intake, that is, it cannot be integrated into a learner's interlanguage, unless the learner consciously attends to the relevant form. In the domain of pronunciation, noticing requires that learners detect the mismatch between their actual phonological output and the target form. In conventional Pakistani classroom settings, this noticing frequently fails to occur because corrective feedback is either absent, delayed, or delivered in whole-class contexts that do not afford learners the opportunity for private reflection and self-correction.

ASR technology operationalizes noticing by providing a closed, real-time feedback loop. Upon producing a segment, learners immediately receive visual and textual information about its accuracy, allowing direct comparison between their production and the target. This mechanism aligns precisely with Schmidt's theoretical requirements for phonological intake. Underwood (2024) extended this perspective by situating ASR within the broader agenda of AI in education, arguing that technologies that facilitate self-monitoring and error detection enact a form of cognitive scaffolding that conventional pedagogy cannot efficiently replicate.

2.2. Krashen’s Affective Filter Hypothesis

Complementary to Schmidt's noticing construct, Krashen’s (1982) Affective Filter Hypothesis posits that anxiety, low motivation, and negative self-concept erect a cognitive barrier that impedes the conversion of comprehensible input into acquisition. In the Pakistani EFL context, this barrier is particularly salient: learners operating within multilingual, socially stratified environments frequently experience speaking anxiety rooted in fear of peer ridicule and public correction (Krashen, 1982). Private ASR practice removes the social audience that triggers this filter, creating conditions under which learners are more willing to attempt phonologically demanding segments and engage in the repetitive practice necessary for phonological automatization.

2.3. CAPT and ASR: Historical Development and Current Evidence

The application of technology to pronunciation training has a substantial history. Early CAPT systems in the 1990s relied on simple waveform visualization and limited phoneme recognition, offering feedback of questionable precision. The emergence of robust ASR engines in the 2000s transformed the field, enabling systems that could reliably detect specific phoneme-level errors and deliver targeted corrective feedback in real time (O'Brien, 2021).

Al-Ghamdi (2022) provided longitudinal evidence from a Saudi Arabian EFL context—one of the few large-class, resource-constrained settings broadly comparable to Pakistan—demonstrating that learners receiving sustained CAPT instruction over an academic year outperformed control participants not only on immediate post-tests but also showed continuing phonological development on measures taken twelve months later.

This finding speaks directly to concerns about the sustainability of ASR-facilitated gains, which the present study addresses through its delayed post-test design.

The most comprehensive quantitative synthesis to date was conducted by Ngo et al. (2024), whose meta-analysis of 15 studies representing 38 effect sizes found an overall medium effect of ASR on ESL/EFL pronunciation (Hedges' g = 0.69). Critically for the present study, Ngo and colleagues' moderator analyses identified explicit corrective feedback as the most effective form of ASR feedback, significantly outperforming indirect feedback modalities such as simple transcription display or binary correct/incorrect signals. Their analysis also identified intermediate proficiency (B1 level) as the proficiency band in which ASR effects are largest, a finding directly relevant to the current participant sample. These evidence-based parameters informed the design of the intervention implemented in the present study.

Lee and Lee (2023) contributed important evidence from a North American CAPT context, examining the ASR feedback functionality of Google Translate as a pronunciation training tool. Their study found that ASR systems demonstrated high accuracy for clearly pronounced target forms but produced more variable recognition for non-standard accented productions, including common L1 phonological transfer errors such as the substitution of /θ/ with /d/ or /t/, a pattern highly prevalent among Pakistani ESL learners. Lee and Lee's finding that explicit error flagging within ASR interfaces substantially increased learners' self-monitoring behaviors aligns with Schmidt's noticing framework and provides a mechanistic account of why explicit ASR feedback outperforms implicit alternatives.

O'Brien (2021) offered a comprehensive review of the trajectory of technology in pronunciation instruction, tracing the transition from discrete-point phoneme training software to integrated CAPT environments and advocating for approaches that address both segmental and suprasegmental features simultaneously. O'Brien argued that the most pedagogically effective CAPT systems are those that provide diverse input, offer clear real-time visual feedback, and establish measurable learning goals, all features present in the ASR intervention used in the current study. O'Brien further noted that the shift toward intelligibility-focused rather than native-speaker-norm-focused instruction has made ASR tools more educationally appropriate, since intelligibility criteria are more attainable for most L2 learners.

Underwood (2024) situated ASR-based pronunciation training within the larger landscape of artificial intelligence in language education, arguing that AI-mediated instruction differs qualitatively from earlier CALL in its capacity for individualized, adaptive feedback at scale. Underwood emphasized that for AI tools to produce durable learning gains, they must engage learners in active cognitive processing rather than passive exposure, a requirement met by ASR systems that prompt explicit error correction and

self-monitoring. This theoretical perspective reinforces the relevance of Schmidt's noticing construct and supports the hypothesis that the active, iterative practice afforded by ASR produces larger and more persistent gains than conventional passive instruction.

Shadish et al.'s (2002) foundational work on quasi-experimental design provides the methodological warrant for the present study. Shadish and colleagues demonstrated that quasi-experimental designs with strong pre-test equivalence, clearly operationalized treatment conditions, and multiple post-test assessment points can approximate the causal validity of randomized controlled trials in applied educational research settings. Their criteria directly informed the present study's design: groups were verified to be equivalent at baseline, the intervention was operationally defined and consistently delivered, and both immediate and delayed post-test assessments were included to evaluate both learning and retention.

2.4. Research Gap: The Pakistani EFL Context

Despite the body of literature reviewed above, a significant research gap remains regarding the use of ASR technologies within the sociolinguistic environment of Pakistan. All major studies to date have been conducted either in East Asian contexts (where L1 phonological transfer patterns differ substantially from those of South Asian learners) or in European and North American settings (where infrastructural constraints differ markedly from Pakistani public universities). No published study has examined the effectiveness of ASR-based pronunciation training among Pakistani ESL learners, a population characterized by:

- Phonological interference from multiple regionally dominant L1s, including Urdu, Punjabi, Sindhi, Pashto, and Balochi, each producing distinct phoneme substitution and prosodic transfer patterns;
- Elevated affective filters arising from the social dynamics of public correction in mixed-proficiency, large-enrollment classrooms; and
- Infrastructural constraints, including limited access to dedicated acoustic environments, unreliable internet connectivity, and low availability of personal computing devices.

The present study is designed to address this gap by generating ecologically valid empirical evidence from within the Pakistani higher education context, thereby contributing to a more geographically and linguistically diverse evidence base for CAPT research.

3. Methodology

This study employs a Quasi-Experimental Mixed-Methods design, consistent with Shadish et al.'s (2002) framework for causal inference in field settings. Quantitative data from pre-tests, immediate post-tests, and delayed post-tests were supplemented with qualitative data from semi-structured interviews, enabling both statistical measurement of

intervention effects and contextual interpretation of learner experiences. All sessions were conducted in participants' natural classroom environment to maximize ecological validity.

3.1. Participants

Forty EFL learners (N = 40) were recruited through purposive sampling at a public-sector university in Lahore, Pakistan. Participants were enrolled in the BS English program (Semesters 4 and 6), held B1 (Intermediate) proficiency on the Common European Framework of Reference (CEFR), and ranged in age from 18 to 22 years. Participants were randomly assigned to either the Experimental Group (n = 20) or the Control Group (n = 20). Full demographic profiles are presented in Table 1.

Table 1: Participant Demographic Profile

Demographic Variable	Experimental Group (n = 20)	Control Group (n = 20)
Age Range	18–22 years	18–22 years
Gender		
Male	10	9
Female	10	11
English Proficiency	B1 (Intermediate)	B1 (Intermediate)
Academic Background	BS English (Sem. 4/6)	BS English (Sem. 4/6)
Ethical Clearances	Informed Consent Obtained	Informed Consent Obtained

Note. All participants provided written informed consent prior to participation. Groups were statistically equivalent on all demographic variables at baseline.

3.2. Measures and Interrater Reliability

Pronunciation was assessed using a standardized reading-aloud and spontaneous speech instrument evaluating both segmental features (individual phonemes) and suprasegmental features (stress, rhythm, and intonation). Responses were scored on a 5-point rubric assessing Intelligibility, Fluency, and Accuracy. To ensure objectivity, two independent raters, both holding postgraduate qualifications in applied linguistics, scored all performances blind to group membership. Interrater agreement was strong (Cohen's $\kappa = .84$), confirming sufficient rating consistency for reliable quantitative analysis.

3.3. Intervention Procedure

The intervention spanned six weeks and comprised eight structured sessions delivered during the first four weeks. The Experimental Group received ASR-based pronunciation instruction with immediate, explicit corrective feedback; the Control Group received equivalent instructional content through conventional teacher-mediated methods with implicit feedback. A two-week gap preceded the delayed post-test to allow assessment of long-term phonological retention. Full intervention details are provided in Table 2.

Table 2: Intervention Protocol

Property	Detail
Total Duration	6 weeks (including 2-week gap before delayed retention test)

Phases	4 distinct phases
Target Groups	Experimental (ASR) and Control (traditional instruction)
Technology Used	Automated Speech Recognition (ASR) software
Assessment Intervals	3 time points: Pre-test (Week 1), Immediate Post-test (Week 4), Delayed Post-test (Week 6)
Feedback Type	Explicit, immediate corrective feedback via ASR vs. implicit teacher feedback

Note. Both groups received equivalent instructional time and content exposure. The independent variable was the modality of corrective feedback.

3.4. Data Analysis Plan

Quantitative data were analyzed using independent samples t-tests to compare group post-test means, paired t-tests to assess within-group change, and one-way ANOVA to compare post-test scores across groups. Practical significance was evaluated using Cohen's d, interpreted according to Cohen's (1988) conventional benchmarks (small = 0.20, medium = 0.50, large = 0.80). Qualitative interview data were analyzed through Braun and Clarke's (2006) six-phase thematic analysis procedure. The full analytical framework is summarized in Table 3.

Table 3: Statistical and Analytical Approach

Property	**Detail**
Design	Mixed-Methods (quantitative + qualitative)
Data Types	Quantitative (test scores) and Qualitative (semi-structured interview data)
Objectives	3 objectives: comparative analysis, effect magnitude, learner perception
Quantitative Tests	Independent samples t-test, one-way ANOVA, Cohen's d
Qualitative Analysis	Thematic analysis via iterative interview coding
Significance Level	$\alpha = .05$ for all inferential tests

Note. $\alpha = .05$ for all inferential tests. Effect sizes interpreted using Cohen (1988) conventions.

4. Results

4.1. Quantitative Analysis: Pre-test and Post-test Scores

Both groups showed improvement from pre-test to post-test; however, the magnitude of improvement differed substantially. The Experimental Group demonstrated a mean gain of 24.25 points (M = 54.25 to M = 78.50), which was highly statistically significant ($t(19) = 12.45$, $p < .001$). The Control Group showed a more modest gain of 7.20 points (M = 55.10 to M = 62.30; $t(19) = 4.12$, $p = .042$). Critically, the two groups were statistically equivalent at baseline ($t(38) = 0.43$, $p = .67$), confirming that post-test

differences are attributable to the intervention rather than pre-existing differences. Full between-group results are presented in Table 4.

Table 4: Statistical Comparison of Pre-test and Post-test Scores

Group	N	Pre-test M (SD)	Post-test M (SD)	t	p
Experimental (ASR)	20	54.25 (6.12)	78.50 (5.40)	12.45	< .001
Control (Traditional)	20	55.10 (5.85)	62.30 (6.25)	4.12	.042

Note. Two-tailed independent samples t-tests. Baseline equivalence confirmed: $t(38) = 0.43$, $p = .67$. Significance threshold $\alpha = .05$.

4.2. Statistical Analysis

4.2.1. One-Way ANOVA

A one-way ANOVA was conducted comparing the post-test scores of the Experimental and Control Groups to assess the overall between-group effect of the intervention. The analysis revealed a statistically significant main effect of group, $F(1, 38) = 48.26$, $p < .001$, confirming that the observed difference in post-test means was not attributable to chance. The ANOVA summary is presented in Table 5.

Table 5: One-Way ANOVA Summary: Post-test Scores by Group

Source	SS	df	MS	F	p
Between Groups	1,850.4	1	1,850.4	48.26	< .001
Within Groups	1,457.8	38	38.36		
Total	3,308.2	39			

Note. $F(1, 38) = 48.26$, $p < .001$. SS = Sum of Squares; df = degrees of freedom; MS = Mean Square. The significant F-ratio confirms that group membership (ASR vs. traditional instruction) was a significant predictor of post-test pronunciation scores.

4.2.2. Effect Size Analysis

To quantify the practical significance of the intervention, Cohen's d was calculated for three key comparisons: (a) the pre-to-post gain within the Experimental Group, (b) the pre-to-post gain within the Control Group, and (c) the between-group difference at post-test. Effect size estimates and their interpretations are summarized in Table 6.

Table 6: Effect Size Estimates (Cohen's d)

Group	Immediate Post-test M	Delayed Post-test M	Mean Decay
Experimental	78.50	76.25	−2.25
Control	62.30	58.15	−4.15

Note. Effect size benchmarks follow Cohen (1988): small = 0.20, medium = 0.50, large = 0.80. The Experimental Group's pre-to-post $d = 1.45$ is classified as very large, indicating that the ASR intervention had a strong and practically meaningful impact on pronunciation accuracy.

The Experimental Group's within-group effect size of $d = 1.45$ substantially exceeds the conventional large-effect threshold of 0.80, indicating that the ASR-based

intervention produced a strong and practically meaningful improvement in pronunciation accuracy. This effect size is notably higher than the overall medium effect (Hedges' g = 0.69) reported in Ngo et al.'s (2024) meta-analysis, though this comparison should be interpreted cautiously given the relatively small sample in the present study. The between-group post-test effect size of d = 1.27 further confirms that the Experimental Group's advantage at the conclusion of the intervention was both statistically significant and large in practical terms.

The Control Group's medium effect size (d = 0.58) indicates that conventional instruction also produced meaningful, though substantially smaller, phonological gains. This is consistent with the general finding in CAPT research that both technology-assisted and conventional instruction improve pronunciation to some degree, but that the additional precision and immediacy of ASR feedback generates incremental gains beyond what conventional methods alone can achieve (O'Brien, 2021; Al-Ghamdi, 2022).

4.3. Long-Term Retention: Delayed Post-test

A delayed post-test administered two weeks after the conclusion of the intervention assessed the durability of pronunciation gains. As shown in Table 7, the Experimental Group retained the majority of its gains, with a mean decay of only 2.25 points (M = 78.50 to M = 76.25), representing a 2.9% decline from the immediate post-test score. The Control Group exhibited a larger decay of 4.15 points (M = 62.30 to M = 58.15), representing a 6.7% decline. A paired t-test confirmed that the difference in decay rates between the two groups was statistically significant, t(38) = 2.89, p = .006, indicating that ASR-based instruction produced more durable phonological learning than conventional instruction.

Table 7: Retention of Pronunciation Gains at Delayed Post-test

Group	Immediate Post-test M	Delayed Post-test M	Mean Decay
Experimental	78.50	76.25	−2.25
Control	62.30	58.15	−4.15

Note. Delayed post-test administered two weeks after conclusion of the intervention (Week 6). Between-group difference in decay rates: t(38) = 2.89, p = .006.

The relatively low decay rate in the Experimental Group is consistent with Shadish et al.'s (2002) prediction that interventions which promote active self-monitoring produce more durable learning than those relying on passive reception of corrective information. The explicit noticing afforded by ASR feedback appears to support deeper phonological encoding, consistent with Schmidt's (1990) account of how noticing facilitates the conversion of input to intake.

4.4. Qualitative Findings: Learner Perceptions

Thematic analysis of the semi-structured interviews with Experimental Group participants yielded three primary themes. Representative participant statements and frequency counts are summarized in Table 8.

Table 8: Qualitative Themes and Representative Participant Statements

Theme	Representative Participant Statement	Frequency
Immediate Corrective Feedback	"The program showed me exactly which sound was wrong so I could fix it right away."	17/20 participants
Reduction in Speaking Anxiety	"I did not feel as nervous as when speaking in front of the whole class, because the computer does not criticize me."	15/20 participants
Technical Interference	"Sometimes the program could not hear me clearly because of noise in the room."	8/20 participants

Note. Frequency counts reflect the number of participants who independently raised each theme during coding. Multiple themes could be identified for a single participant.

4.4.1. Immediate Corrective Feedback

Seventeen of the twenty participants reported that the ability to self-correct in real time enabled them to master phonologically challenging contrasts, particularly /θ/ versus /d/, /v/ versus /w/, and the fricative /ʒ/. Participants described the feedback as more actionable than verbal corrections received in class, noting that the visual salience of the ASR output made the nature of the error clear in a way that oral teacher correction rarely achieved. This theme aligns with O'Brien's (2021) finding that clear real-time feedback with visual aids is among the most critical features of effective CAPT systems.

4.4.2. Reduction in Speaking Anxiety.

Fifteen of twenty participants described a significant decrease in speaking anxiety when using ASR. Participants attributed this primarily to the absence of peer observation and the non-judgmental nature of the computer system. One participant stated: "I did not feel as nervous as when speaking in front of the whole class, because the computer does not criticize me." This finding is consistent with Krashen's (1982) Affective Filter Hypothesis and corroborates Lee and Lee's (2023) observation that ASR environments reduce the social evaluation threat that suppresses learners' willingness to attempt phonologically risky utterances.

4.4.3. Technical Interference.

Eight participants noted that background noise in shared classroom environments occasionally disrupted accurate speech recognition, producing erroneous feedback that undermined their confidence in the system. This concern echoes the technical limitation identified by Lee and Lee (2023), who found that ASR accuracy for L2 speakers is sensitive to acoustic environment quality, and points to an infrastructural challenge specific to the resource-constrained Pakistani university context.

5. Discussion

The results of this study provide strong empirical support for the effectiveness of ASR-based feedback in improving the pronunciation accuracy of Pakistani ESL learners. The Experimental Group's substantially greater gains, both immediately post-intervention and at the delayed retention test, demonstrate that computer-mediated explicit feedback is more effective than teacher-mediated implicit feedback in the higher education context examined here. The large effect sizes obtained (d = 1.45 within-group; d = 1.27 between-group) compare favorably with the existing literature and warrant detailed discussion in relation to the theoretical framework and prior empirical work.

5.1. Theoretical Relevance: Noticing and Intake

The findings are consistent with Schmidt's (1990) Noticing Hypothesis. The immediate, visually salient feedback provided by ASR software enabled learners to notice the gap between their interlanguage phonology and target norms in real time—a condition Schmidt identified as necessary for the conversion of input to intake. In conventional classroom settings, pronunciation feedback is typically delayed, generalized across multiple learners, and rarely experienced as personally relevant by individual students. By contrast, the ASR feedback loop confronted each learner with an individualized, phoneme-specific signal at the precise moment of production, maximizing the conditions for Schmidt's noticing to occur.

This interpretation extends Underwood's (2024) argument that AI-mediated language instruction engages active cognitive processing in ways that passive conventional instruction cannot replicate. The active self-monitoring and iterative correction enabled by ASR transformed pronunciation practice from a passive reception exercise into an active form-focused production cycle, consistent with Schmidt's emphasis on attention and awareness as prerequisites for acquisition.

5.2. Affective Dimensions: Anxiety Reduction and the Filter

The qualitative evidence for reduced speaking anxiety constitutes an important complementary finding. Krashen's (1982) Affective Filter Hypothesis predicts that learners in low-anxiety, non-threatening conditions are better able to process and acquire the linguistic input they receive. The private, non-evaluative nature of ASR practice appears to have lowered the affective filter sufficiently for learners to engage willingly with phonologically challenging targets that they would avoid in classroom settings. This mechanism may partially account for the large quantitative gains observed: not only did ASR provide better feedback quality, but it also increased the quantity and risk-tolerance of practice itself.

This finding resonates with Lee and Lee's (2023) observation that ASR environments remove the social evaluation threat inherent in classroom pronunciation activities. In the Pakistani sociocultural context, where public academic failure carries significant social stigma, this affective dimension may be particularly consequential. The

magnitude of anxiety reduction reported by participants, 15 of 20 raising the theme unprompted in interviews, suggests that the affective benefits of ASR are not merely incidental but represent a core mechanism through which ASR enhances pronunciation development in this population.

5.3. Comparison with Prior Research

The present study's findings extend and contextualize the international literature in several important ways. The large between-group effect size (d = 1.27) substantially exceeds the overall effect size reported in Ngo et al.'s (2024) meta-analysis (Hedges' g = 0.69), suggesting that the combination of explicit corrective feedback and an intermediate-proficiency B1 sample—both identified as effect-size moderators by Ngo and colleagues, may be particularly favorable conditions for ASR-based pronunciation learning. This interpretation is supported by Ngo et al.'s finding that ASR with explicit feedback is "largely effective" and that benefits are greatest for intermediate learners, precisely the profile of the current participant sample.

This study also corroborates Al-Ghamdi's (2022) longitudinal evidence from a comparable large-class EFL context in Saudi Arabia, demonstrating that the infrastructure constraints characteristic of public-sector universities in Muslim-majority South and Southwest Asian countries do not preclude meaningful ASR-facilitated gains. Al-Ghamdi found that sustained CAPT instruction produced continuing phonological development beyond the training period; the present study's delayed post-test, while covering only a two-week interval, provides initial evidence that the gains produced by ASR persist beyond immediate instruction in the Pakistani context.

The present results are consistent with O'Brien's (2021) systematic review, which concluded that CAPT systems offering real-time visual feedback and explicit error identification produce the largest and most durable pronunciation gains. O'Brien's observation that the field has shifted toward intelligibility-focused goals is particularly relevant to the Pakistani context, where the primary communicative objective is mutual intelligibility rather than native-speaker approximation, a goal for which ASR tools, calibrated against intelligible L2 speech, are well suited.

Shadish et al.'s (2002) methodological framework is also vindicated by the present findings. The quasi-experimental design, anchored by verified baseline equivalence and augmented with a delayed post-test, produced causal evidence of adequate internal validity despite the absence of full randomization. The use of an intent-to-treat analytic approach and Cohen's d to quantify practical significance further strengthens the evidential contribution of the study within the limits of its sample size.

5.4. Mitigating Limitations

Several limitations constrain the generalizability of these findings. The sample is restricted to one institution in Lahore and does not capture the full phonological diversity

of Pakistani ESL learners: learners whose primary L1 is Sindhi, Pashto, Balochi, or Punjabi may respond differently to ASR feedback than the predominantly Urdu-dominant sample recruited here. The four-week intervention, while sufficient to produce statistically significant and large-effect gains, does not permit conclusions about phonological development beyond this period; the observed decay in the delayed post-test, modest though it is, indicates that sustained practice is required to prevent deterioration, consistent with Al-Ghamdi's (2022) longitudinal findings. Technical issues related to acoustic environment quality, raised by 8 of 20 participants, represent a context-specific limitation that future studies should address by including decibel-level measurements and acoustic treatment as design variables.

6. Conclusion

This study provides compelling evidence that ASR-based pronunciation feedback significantly outperforms conventional teacher-mediated instruction in improving the phonological accuracy of Pakistani ESL learners. Experimental participants demonstrated substantially greater gains in segmental and suprasegmental features ($d = 1.45$), superior long-term phonological retention (mean decay -2.25 vs. -4.15), and markedly reduced speaking anxiety. One-way ANOVA confirmed the statistical reliability of the between-group difference ($F(1, 38) = 48.26$, $p < .001$), and the effect sizes observed are large by any conventional benchmark.

Theoretically, these outcomes are coherent with both Schmidt's (1990) Noticing Hypothesis, in that ASR makes phonological error perceptually salient at the moment of production, and Krashen's (1982) Affective Filter Hypothesis, in that the private, non-evaluative practice environment reduces the anxiety that impedes pronunciation development. Practically, the findings demonstrate the viability of ASR-enhanced pronunciation instruction within the infrastructural realities of Pakistani public higher education.

6.1. Pedagogical Implications

These findings carry several implications for English language instruction in Pakistan:

1. **Blended Learning Integration.** Institutions should incorporate ASR-based pronunciation tools as a structured complement to classroom instruction, particularly in high-enrollment programs where individualized teacher feedback is impractical.
2. **Teacher Support and Workload.** ASR tools can significantly reduce the burden on teachers delivering feedback across large classes, freeing instructional time for higher-order communicative activities and metalinguistic discussion.

3. **Learner Autonomy.** The private, self-paced nature of ASR practice fosters autonomous learning habits, encouraging learners to take ownership of their phonological development beyond formal instructional hours.

7. Recommendations for Future Research

Future studies should expand the sample to include learners from multiple provinces and diverse L1 backgrounds, particularly Sindhi, Pashto, and Balochi speakers, to examine differential effects across distinct phonological transfer patterns. Longitudinal designs spanning a full academic year would clarify the sustainability of gains and identify optimal intervention durations. Studies manipulating acoustic environment quality would help disentangle the technical and pedagogical contributions to ASR effectiveness. Finally, mixed-methods research examining teacher perspectives on ASR integration would complement the learner-centered evidence generated here and inform the design of professional development programs for Pakistani EFL educators.

References

Al-Ghamdi, N. (2022). Longitudinal study of CAPT in EFL pronunciation instruction. *Journal of Computer Assisted Learning.* https://doi.org/10.1111/jcal.12665

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.

Krashen, S. D. (1982). *Principles and practice in second language acquisition.*Pergamon Press. http://www.sdkrashen.com/content/books/principles_and_practice.pdf

Lee, J., & Lee, S. (2023). Evaluating ASR feedback for L2 pronunciation: A focus on common error types. *Language Learning & Technology.* https://hdl.handle.net/10125/77382

Ngo, T. T.-N., Chen, H. H.-J., & Lai, K. K.-W. (2024). The effectiveness of automatic speech recognition in ESL/EFL pronunciation: A meta-analysis. *ReCALL, 36*(1), 4–21. https://doi.org/10.1017/S0958344023000113

O'Brien, M. G. (2021). Teaching and learning pronunciation with technology: Current possibilities, pedagogical recommendations, and directions for the future. *The Modern Language Journal, 105*(1), 232–251. https://doi.org/10.1111/modl.12689

Schmidt, R. W. (1990). The role of consciousness in second language learning. *Applied Linguistics, 11*(2), 129–158. https://doi.org/10.1093/applin/11.2.129

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Houghton Mifflin.

Underwood, J. (2024). *Artificial intelligence in language education.* Routledge. https://doi.org/10.4324/9781003362142