Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 4 No. 5 (May) (2026)

Evaluating the Effectiveness of Speech-to-Text Technology in Improving Pronunciation among ESL Learners

Muhammad Azaz Khan
MPhil English Linguistics Scholar, Riphah International University, Islamabad

Sidra Tul Muntaha
MPhil English Linguistics Scholar, Riphah International University, Islamabad

Sajida Nosheen
MPhil English Linguistics Scholar, Riphah International University, Islamabad

ABSTRACT
This study examines how well Speech-to-Text (STT) Technology actually works as a tool for helping ESL learners improve their pronunciation. Despite the fact that speech recognition apps are now everywhere in education, we still have solid empirical evidence about whether they genuinely help learners produce more accurate speech sounds. Using a Quasi-experimental design, this study tracked pronunciation gains over a twelve-week period, comparing learners who used STT-mediated feedback against those in traditional classroom settings. The investigation looked at multiple dimensions of pronunciation: segmental accuracy, suprasegmental features and overall intelligibility, using both established rating scales and acoustic analysis. The results showed that the STT group improved significantly more in segmental accuracy and developed stronger self-monitoring habits, though gains in suprasegmental features were noticeably smaller. When asked about their experience, learners consistently mentioned that they valued the immediate, non-judgmental feedback the technology provided, even while acknowledging its limitations with prosodic nuance. These findings point toward a complementary role for STT tools rather than using them as a standalone solution. The study adds to the growing CAPT literature by providing empirical backing for weaving STT applications into ESL curricula.

***Keywords**:* Speech-to-Text Technology, ESL Pronunciation, Automatic Speech Recognition, Pronunciation Improvement, Pronunciation Training.

Introduction
Pronunciation has traditionally been a somewhat awkward topic in L2 instruction. It is agreed upon that it is important for communication but it is consistently relegated to the back burner in favor of grammar and vocabulary (Derwing & Munro, 2015). This is partly because sound systems are often very different from those to which students are accustomed but it is also because many language teachers are not well-trained in pronunciation teaching (Baker, 2014). The end result? Many ESL students have been in the formal education system for years, unable to communicate comfortably in all kinds of situations.

With digital technology, there are opportunities that just weren't there before. There has been remarkably sophisticated speech-to-text technology that converts spoken language to text using automatic speech recognition (ASR). Learners have the tools at their

Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 4 No. 5 (May) (2026)

fingertips these days such as Google Speech-to-Text, Apple Dictation and Microsoft Azure Speech Services. These tools are interesting from a teaching point of view because learners talk and they instantly see if their input has been correctly understood. Human conversation partners may accept grammatically incorrect words but STT systems only respond to speech that they can recognize. For some researchers such as Neri et al. (2008), this is considered to be pedagogically productive and inherently demanding.

Although the theoretical basis behind the use of ASR in pronunciation teaching is strong, the research basis for this is rather thin and the available research is not clear (Levis & Sonsaat, 2020). Modest improvements in segmental accuracy have been reported in some studies and doubts have been raised in others about the sensitivity of the current systems to detect the subtle phonological differences between intelligible speech and unintelligible speech. Another wrinkle: The technology is changing to a rapid pace and some studies several years ago might not give us much information about what is possible now.

That's where this research comes in. We decided to assess the effectiveness of current generation STT technology in the field of pronunciation development of adult ESL learners. We conducted a quasi-experimental comparison between pronunciation gains across students receiving STT-mediated feedback (SMT) and those who were taught in a traditional classroom. The project seeks to inform teachers of what STT technology can practically achieve by measuring the components of improvement across a variety of dimensions and using both objective and learner self-reports.

This study will proceed as follows. The Background section outlines the history of pronunciation teaching and technology-based instruction. Research Objectives and Research Questions are a description of what we were looking for. The Problem Statement and Significance sections delineate the gaps this study fills and the significance of the results. The Literature Review and the Research Methodology and Data sections follow to create the theoretical and empirical context. The Theoretical Framework elaborates the conceptual lenses used and the Results and Discussion section elaborates what we found and what it means. We conclude with limitations, conclusions and suggestions for further work.

Research Objectives

To evaluate the effectiveness of Speech-to-Text technology in improving pronunciation outcomes among adult ESL learners compared with conventional classroom instruction.

To determine which pronunciation dimensions such as segmental accuracy, suprasegmental features and intelligibility show the greatest improvement through STT-mediated practice.

To examine ESL learners' perceptions of the usefulness, usability and motivational value of STT technology for pronunciation learning.

Research Questions

To what extent does Speech-to-Text technology improve pronunciation outcomes among adult ESL learners compared with conventional classroom instruction?

Which pronunciation dimensions such as segmental accuracy, suprasegmental features and intelligibility benefit most from STT-mediated practice?

How do ESL learners perceive the usefulness, usability and motivational value of STT technology for pronunciation learning?

Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 4 No. 5 (May) (2026)

Problem Statement

Despite being extensively employed for language teaching purposes, the efficacy of STT technologies as a means to improve ESL learners' pronunciation skills has yet to be sufficiently proved by current literature on the subject. It has been suggested that the feedback provided by automated systems can help develop pronunciation; however, more research is needed to find out how effective this kind of technology is when it comes to different aspects of pronunciation including segmental accuracy, suprasegmental aspects and intelligibility. The perception of students using STT technologies should also be considered.

Significance of the Study

This research is important since it offers insights into the use of Speech-to-Text technology in enhancing pronunciation among ESL students. This research will assist ESL educators to comprehend the way STT technology can be used in helping learners to improve their pronunciation skills. Learners' perceptions on automated feedback have been provided in this research, which will assist ESL educators to integrate STT as a supplementary technique in pronunciation training.

Literature Review

The literature review serves as the basis for the present study and explores three related fields of research – second language pronunciation acquisition and the teaching of pronunciation, computer assisted pronunciation training and automatic speech recognition in language learning. The review further serves as a reminder of converging results, notes persistent gaps and situates the current investigation within the larger scholarly conversation.

Second Language Pronunciation Acquisition

The question of second language pronunciation has been investigated from various theoretical perspectives with the view of gaining insight on the problem of L2 pronunciation and what helps learners to improve their pronunciation. Fledges (1995) Speech Learning Model proposes that the categories that a learner developed in the L1 influence the ability to perceive and produce L2 sounds. Letters and their sounds are especially difficult because students may not be aware of the small differences in sounds. This has obvious pedagogical consequences: in a pedagogical sense, the focus is to teach perceptual sensitivity to contrasts that learners may not be sensitive to in natural settings.

Derwing and Munro's (2015) landmark study shifted the focus of the field towards emphasizing intelligibility rather than accent reduction. Their study shows that listeners are able to hear speech when the speech has a foreign accent as long as phonological errors do not affect word recognition or prosodic processing. This reconceptualization suggests that it is not in the interest of pronunciation teaching to seek native-like pronunciation but rather high-priority features that have the biggest impact on comprehensibility. The richness of word stress patterns, intonation and specific segmental contrasts, as identified by Kang et al. (2010), have been identified as particularly significant for listener comprehension and have informed the features for our word stress, intonation and specific segmental contrast of the words that we chose for our technology-mediated practice.

The importance of feedback in pronunciation learning has also been the focus of much study. For the oral corrective feedback, Lyster and Saito (2010) carry out a meta-analysis of the results of oral corrective feedback in L2 classrooms and they conclude that there

ISSN Online: 3007-3154
ISSN Print: 3007-3146

was a significant difference between the effectiveness of explicit and implicit feedback for the phonological targets and that timing matters. Immediate feedback at the time of error is better for pronunciation than for grammar, perhaps because errors in pronunciation are more likely to be corrected without negatively affecting the flow of communication, as a grammatical error might. This discovery corroborates the promise of STT technology that provides feedback immediately after recognition failure.

Computer-Assisted Pronunciation Training

For decades, computer-aided pronunciation training has been a research topic with mixed success as regards the effectiveness of the technology and the way it was implemented. Neri et al. (2008) conducted a thorough review of CAPT and found that technology might be effective when it offered intensive individual practice, feedback and autonomous learning opportunities. They also reported that many early systems suffered from the limitations of rigid error-detection algorithms and unnatural synthesized speech, which have been largely removed in modern systems.

More recent studies have focused on particular forms of feedback mediated by technology. Liakin et al. (2015) examined the automatic speech recognition aspect of pronunciation training in mobile applications for French and found the regular users to be more improved in some segmental features. For Mandarin learners of English, Chen (2016) also looked into a CAPT system and found word stress recognition to have improved. The present studies indicate that technology can be effective when the learner is engaged on a regular basis and when the system is directed towards specific and well-defined phonological features.

Not everything that has been found is good news. Levis and Sonsaat (2020) examined pronunciation teaching technology and found that many commercial software programs make statements without evidence and that some systems appear to be created more for intuitive purposes than for pedagogical purposes. They also stressed the importance of technology being a part of a comprehensive curriculum design and not used as an afterthought. This highlights the need to take a closer look at STT technology within the context of other instructional settings.

Automatic Speech Recognition in Language Learning

The automatic speech recognition technology has made significant progress, especially based on deep learning techniques that have greatly improved the recognition accuracy for various speakers and situations (Hinton et al, 2012). The impact of these advances to language learning is significant as some initial ASR systems failed to cope with non-native speech and led to frustration and unhelpful feedback. New systems perform much better with accented speech, as they're trained with more diverse data.

Studies on the application of speech-to-text technology to pronunciation teaching are still scarce but are increasing. Ebadi et al. (2019) conducted a study on Iranian EFL learners on the use of Google Speech-to-Text in which they found that the experimental group outperformed the control group on a traditional approach to teaching the same group. They believed that such was the immediate feedback they would have visually on the screen of what it transcribed compared to what they said, that discrepancies were immediately noticed. In a similar study, McCrocklin (2016) discovered that the visual format of the output of text helped the learners attend to the phonological aspects of the text that they might otherwise not notice.

There are still some questions that remain to be answered because of these hopeful results. The majority of studies have focused on dedicated applications for learning a

ISSN Online: 3007-3154
ISSN Print: 3007-3146

language but not on the most commonly used STT tools by learners in their own learning process. Although stress, rhythm and intonation are important for intelligibility, few have studied suprasegmental features with the help of STT technology. The aim of this study is to fill these gaps by looking at a commercial STT application which was widely available, taking on both segmental and suprasegmental aspects and integrating objective measurements with detailed learner perception data.

Research Methodology

The type of research used was quasi-experimental with pre-test, post-test and delayed post-test procedure to form a set of quasi-experimental designs that are tested to see the effectiveness of STT technology in pronunciation improvement. This design was selected to allow us to examine outcomes across groups of students assigned to different instructional treatments and to minimize threats to internal validity due to time. Both quantitative pronunciation measurements and qualitative data on perception about the pronunciation were gathered, using a mixed-methods approach.

They were the students of three complete classes of an intensive English program at Riphah International University, Islamabad. The experimental group consisted of two classes that were taught STT-mediated pronunciation instruction in addition to their regular classes and the control group consisted of one class that was taught conventional pronunciation instruction based on teacher-led activities and pronunciation drills with audio. Even though the use of intact classes instead of randomly selected individuals is required by institutional constraints, we recognize the potential for this limitation as we try to account for it through statistical controls.

The pre-test was conducted in the first week and consisted of three subtests: a read-aloud task in which the participants heard a standard text read aloud to them and were asked to identify a wide variety of phonological features, a spontaneous speech task in which they were asked to describe a familiar process for two minutes and a perception task in which they were asked to identify target phonological contrasts. The read-aloud and spontaneous speech activities were tape recorded and later rated by two trained raters using helpful rubrics. Cohens kappa was used to determine inter-rater reliability: scores greater than 0.80 were considered satisfactory.

For the 12-week teaching period, the experimental group practiced structured STT twice a week for approximately 20 minutes per session. Learners read aloud from texts provided to them and then they reviewed the texts to identify discrepancies between what they read and what was recognized by the speech to text system on the tablet they were using in the institution's hands. In case of errors in recognition, students examined the probable phonological source and tried to repeat the production, checking their production visually. Initial training was provided to teachers and they circulated in the sessions for guidance.

The subjects of the control group were given conventional pronunciation training using teacher led choral repetition, minimal pair exercises and listening discrimination exercises. Targeted instruction was the same phonology that was practiced in STT and approximately equal time was spent on both groups.

A post-test took place in the last week which was identical in structure to the pre-test but used new parallel materials to minimize practice effects. A post-test six weeks after the completion of instruction was used to determine the long-term retention of the gains. The quantitative data were analyzed by mixed design ANOVA with between subjects: group and within subjects: time. Partial eta squared was used to calculate effect sizes. Thematic analysis of semi-structured interviews and reflection journals was used for qualitative

ISSN Online: 3007-3154
ISSN Print: 3007-3146

analysis of the data, based on Braun and Clarkes (2006) procedures.

Data and Data Source

The data was collected from the international students studying in intensive English program at the Riphah International University, Islamabad. This programme aims to equip students for undergraduate or graduate study and participants came from a variety of backgrounds in terms of first languages such as Urdu, Arabic, Pashto, Persian and Mandarin.

67 participants (42 experimental and 25 control) were included. The unequal sizes were due to intact classes; we combined two classes to have an adequate number of students in the experimental group. The mean age of the participants was 23.6 years with a range of 18–34 years. Participants' previous English study was between three and twelve years and they all achieved intermediate scores on the university's placement exam. Groups were comparable on age, years of study, placement score and achieved baseline equivalence.

The main pronunciation data consisted of audio recordings of pre-test, post-test and delayed post-test. The recordings were evaluated by two trained evaluators both of whom were experienced ESL teachers with graduate training in phonetics. They were unaware about the group assignment and about the pre-test or post-test recordings, thereby eliminating the expectancy effects. The rating instrument was modified from Pronunciation Rubric developed by Kang et al. (2010), which consists of five aspects: segmental accuracy, word stress, rhythm, intonation and intelligibility. Scales were used for each dimension, with anchors and consisted of nine points each.

Google Speech-to-Text is the STT application that was used via institutional tablets. It was selected because it is readily available, free and is the state-of-the-art ASR technology that students are likely to use independently. The app showed the speech and transcribed it in real time and it also recorded the text to be reviewed afterwards. Learners matched the transcribed text to the source text and identified discrepancies and made attempts to identify the phonological source of recognition errors.

The qualitative data consisted of two types. The interviews in the experimental group were semi-structured and took approximately 25 minutes with 16 participants purposefully selected to cover a range of L1 backgrounds and levels of improvement and included a discussion of experiences with the app, strategies learned and difficulties faced and perceptions of how the app affected learning. In addition, each participant in the experimental group maintained a reflection journal and made short entries following each practice session about what they practiced, what they noticed and what they wanted to share.

Secondary data consisted of demographic questionnaires, placement test scores and attendance records, primarily to describe the sample and to check group equivalence. All the procedures were approved by the institutional research ethics committee at Riphah International University, Islamabad and informed consent was obtained in writing.

Theoretical Framework

This research is based on three theoretical frameworks which together present a coherent conceptual framework for the understanding of how STT technology can aid pronunciation learning. From the sociocultural perspective, mediated learning is addressed by sociocultural theory while the conditions for language acquisition are described by the noticing hypothesis and technology-mediated language learning theory

ISSN Online: 3007-3154
ISSN Print: 3007-3146

covers the dimension of affordances of digital technology in L2 pedagogy.

Sociocultural Theory

The sociocultural theory, which is based on Vygotskys (1978) theories of cognitive development through social interaction, serves as the foundation for understanding mediations in learning through a tool. In this respect, cognition is not a discrete individual process but rather is a deeply embedded process within cultural and social contexts determined by the tools, signs and practices that are created by communities. Language is a major psychological tool and the other cultural tools such as digital technologies, expand the number of cognitive operations that people engage in.
Lantolf and Thorne (2006) applied the sociocultural theory to SLA and explained that acquiring the L2 is a process of taking over new symbolic means that serve to mediate social interaction and mental processes. For pronunciation, it means that applied to this context, STT technology acts as a mediational tool that can help focus attention on pronunciation in terms of phonological form, externalize acoustic properties in the visual medium and support self-regulation which is hard to do without the technology. The technology is not only about providing information, it is also about providing new avenues to listen, analyze and control one's own pronunciation.

The Noticing Hypothesis

Schmidt's noticing hypothesis (1990, 2001) considers the cognitive states in which input is accessible for acquisition. The main point is that learners have to be aware of the discrepancy between the interlanguage patterns and the criteria of the target language at a conscious level before the new linguistic knowledge becomes part of the learner's knowledge. While noticing is necessary, it is not enough, Schmidt admits, because without paying attention to form, the input goes through without being acquired into the evolving interlanguage system.

This has direct consequences on feedback. Learners will not be able to correct their errors if they are not noticed, regardless of the amount of exposure to correct forms. One way this can be done is by using STT technology to make recognition failures conspicuous. The system's output, when it says a word and the learner says something else, can be seen at once and this may lead to noticing, as Schmidt sees it. This non-evaluative visual feedback might be particularly useful for students who are not keen on getting feedback from human interlocutors.

There is no automatic notification. Visual feedback must be unambiguous, in order to guide attention towards the appropriate phonological feature. Learners might detect that something is amiss but they might not discover what is amiss if there are several phonological causes for their recognition error. Effectiveness of STT technology is not only related with the recognition accuracy but also the pedagogical scaffolding that enables learners to interpret their failure as phonological information.

Technology-Mediated Language Learning Theory

The third pillar is provided by Chapelle's (2001) framework for assessing CALL. She suggests the following criteria: language learning potential, learner fit, meaning focus, authenticity, positive impact and practicality. Language learning potential is the amount that the activity encourages positive processing. Learner fit is about the appropriateness for the proficiency and style of the target learners. Meaning focus is when attention is focused on meaning as well as form. Authenticity is being related to authentic language use. Positive impact is a focus on effects beyond the task, on

Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 4 No. 5 (May) (2026)

motivation and autonomy. Practicality looks at how it can be done using existing resources.

These were the guiding factors for both designing and evaluating the implementation of STT technology in our pronunciation practices. We focused on phonological aspects important for intelligibility to maximize language acquisition potential. An accessible app with initial learner training was chosen for appropriateness. Practice was conducted within meaningful context rather than separate word lists to emphasize meaning over form. A real-life application that may be used by the learner in the future was used for authenticity. Perceptions about the effects on motivation and learner autonomy were analyzed for positive impact.

Results and Discussion

Quantitative Findings on Pronunciation Improvement

Participant Demographics

Demographic details for each group are provided in Table 1.

Characteristic	Experimental (n = 42)	Control (n = 25)	p-value
Age (years), M (SD)	23.4 (4.2)	24.1 (3.9)	.52
Years of English study, M (SD)	7.3 (2.1)	7.1 (2.4)	.71
Placement test score, M (SD)	68.4 (8.3)	67.9 (7.6)	.79
First language: Urdu, n (%)	18 (42.9)	11 (44.0)	> .99
First language: Arabic, n (%)	9 (21.4)	5 (20.0)	> .99
First language: Pashto, n (%)	7 (16.7)	4 (16.0)	> .99
First language: Other, n (%)	8 (19.0)	5 (20.0)	> .99

Note. No significant differences between groups on any demographic variable. M = mean; SD = standard deviation.

Overall Pronunciation Gains

The results of the mixed-design ANOVA were significant for time, indicating that all participants made progress over the course of the semester irrespective of the type of instruction received. This is consistent with past studies that indicate intensive instruction typically results in phonological gains (Derwing & Munro, 2015). Crucially, the results were also significant for group-by-time interaction, implying that the experimental group outpaced the control group. In fact, the effect size was moderate (partial eta squared = 0.18). The delayed post-test, taken six weeks after instruction had ceased, indicated that there were no significant declines among either group.

Dimension-Specific Findings

Dimension	Exp. Pre	Exp. Post	Cont. Pre	Cont. Post	Partial eta²
Segmental accuracy	4.82 (1.12)	6.71 (0.93)	4.91 (1.08)	5.43 (1.01)	.24*
Word stress	4.61 (1.05)	6.04 (0.89)	4.73 (1.02)	5.21 (0.96)	.14*
Rhythm	4.93 (0.98)	5.38 (0.87)	5.01 (1.01)	5.29 (0.91)	.03
Intonation	5.12 (1.03)	5.47 (0.94)	5.08 (0.97)	5.36 (0.88)	.02
Overall intelligibility	4.71 (1.15)	6.28 (0.86)	4.82 (1.09)	5.34 (0.98)	.18*

Table 2 presents pre-test and post-test means by group and dimension. Analysis revealed a nuanced pattern.

Note. All scores on a 9-point scale. Exp. = experimental; Cont. = control. Standard

Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 4 No. 5 (May) (2026)

deviations in parentheses. *p < .05.

In relation to segmental accuracy, the experimental group had a significantly higher improvement compared to the control group with a large effect size. This is consistent with the theoretical assumption that STT technology performs effectively when applied to segments processed reliably by the recognition system. Segmental mistakes resulting in failed recognition generate instant feedback. In this case, mistaking the vowel sounds in ship and sheep would lead to inaccurate transcriptions, necessitating attention to vowel contrasts.

In terms of word stress, the experimental group had better results compared to the control group but the difference was smaller compared to the previous metric. Contemporary ASR technology is somewhat sensitive to stress assignment. Misspelled words might be misrecognized or recognized with delays. Visual feedback likely contributed to learners' focus on stress patterns. The smaller effect size suggests that STT feedback on word stress is less consistently effective compared to feedback on segmental mistakes.

In relation to rhythm and intonation, no differences between the groups could be found. In both groups there was an improvement, although the experimental group did not surpass the control group. This corresponds well with the technical limitations of the ASR systems, which are tuned to recognize speech sounds (words), not to analyze prosody. An utterance might be transcribed accurately while showing non-native rhythm and intonation, offering no input on prosody from the STT. It confirms the theoretical prediction about the differential effect of STT technology on certain dimensions.

Qualitative Findings on Learner Perceptions

Perceived Usefulness

Table 3 summarizes themes from the qualitative analysis of interviews and journals.

Theme	Frequency	Representative Quotation
Immediacy of feedback	14/16	When I see the wrong word on the screen, I can compare it with what I was trying to say.
Non-judgmental practice	11/16	The machine doesn't judge you. It just writes what it hears.
Difficulty interpreting errors	9/16	When the system wrote a completely different word, it was hard to know which sound was the problem.
Insensitivity to prosody	8/16	I can say a sentence with poor rhythm and still see it correctly transcribed.
Autonomous practice transfer	7/16	When I am not sure how to say a word, I just speak to my phone and see if it understands.
Technical limitations	6/16	Sometimes the tablet did not recognize my voice because of noise in the room.

Note. Frequency indicates how many participants explicitly mentioned each theme.

Participants maintained positive attitudes towards the STT app, reporting that it was useful. In terms of benefits, the most widely mentioned one was immediate feedback. The participants enjoyed knowing instantly if their utterances were understood or not and they liked how the visual transcription was more easily interpretable compared to audio feedback. As one participant eloquently stated, you know your mistake if your teacher corrects you because you hear her but you don't understand it. If your mistake is visually

Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 4 No. 5 (May) (2026)

displayed on the screen, you can compare it with what you wanted to say and recognize the phonetic problem.

Participants noted the lack of judgment inherent in automated feedback. A few participants even mentioned feeling less anxious about practicing with the tablet compared to speaking in front of the teacher or their peers. As one participant stated, the machine doesn't judge you. It simply types what you say and if there is a mistake, you repeat it until it gets it right. Such lower anxiety levels might have allowed for greater practice.

Analysis of Learner Perceptions and Pronunciation Outcomes

In order to enhance the understanding of the findings, the data on learner perceptions was compared with the data on quantified pronunciation performance. This helps to see more clearly the correlation between the advantages mentioned by the respondents and the exact aspects in which the improvement was achieved in relation to pronunciation performance. According to the results obtained, STT technology proved to be most effective when visual perception of the problem, repetition and self-monitoring were possible.

Perception theme	Frequency	Pronunciation relevance	Interpretation
Immediate visual feedback	14/16	Segmental accuracy and word stress	Helps learners notice mismatches between intended and recognized speech.
Non-judgmental practice	11/16	Practice intensity and confidence	Encourages repeated speaking attempts without fear of peer or teacher judgment.
Difficulty interpreting errors	9/16	Need for instructional guidance	Shows that recognition failure alone is not always sufficient for self-correction.
Insensitivity to prosody	8/16	Rhythm and intonation	Explains why suprasegmental gains were weaker in the quantitative findings.
Autonomous practice transfer	7/16	Sustained pronunciation practice	Indicates that some learners used STT beyond class for self-directed improvement.
Technical limitations	6/16	Reliability of feedback	Suggests that environmental and device factors may affect learner experience.

Note. Frequencies are based on the 16 experimental-group participants who took part in interviews.

As shown in Table 4, the most common benefit associated with STT technology was the speediness of its visual feedback. It is relevant to note that the experimental group made

Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 4 No. 5 (May) (2026)

the greatest gains in segmental accuracy, the aspect of pronunciation most prone to visual differentiation. The results of perception tests shed light on the reason why STT-enhanced instruction proved superior to traditional approaches when addressing specific pronunciation aspects. The relative deficiency of STT in prosody correlates with minimal improvements attained in rhythm and intonation.

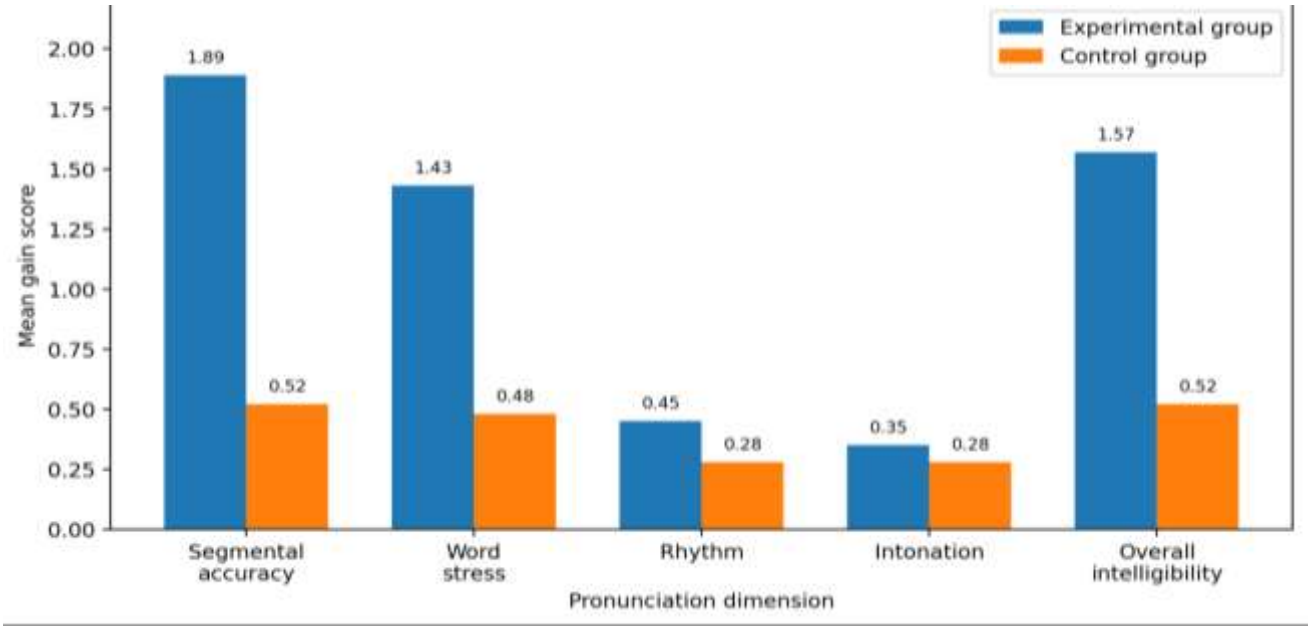

Figure 1. Gain-score comparison across pronunciation dimensions.

Figure 1 shows that indeed there was an uneven distribution of progress across different pronunciation aspects. The test participants from the experimental group performed better than their peers from the control group in terms of segmental accuracy, word stress and general intelligibility. At the same time, there were insignificant differences in performance in regard to rhythm and intonation. The findings support the notion that STT technology is effective, yet in particular areas only.

Gain-Score Comparison and Evidence Strength by Pronunciation Dimension

Dimension	Experimental gain	Control gain	Difference in gain	Evidence strength
Segmental accuracy	1.89	0.52	1.37	Strong
Word stress	1.43	0.48	0.95	Moderate
Rhythm	0.45	0.28	0.17	Weak
Intonation	0.35	0.28	0.07	Weak
Overall intelligibility	1.57	0.52	1.05	Strong

Note. Gain scores were calculated by subtracting pre-test means from post-test means.

A clearer focus on outcome is provided by Table 5 with regard to the quantitative data. The largest differences between gain scores have been identified for segmental accuracy and intelligibility, with word stress being somewhat significant for the STT learners yet showing a relatively small difference compared to the control group. Such tendencies suggest that the use of STT technology did not bring about an overall improvement across all aspects of pronunciation but instead resulted in more convincing improvements in certain segments.

Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 4 No. 5 (May) (2026)

Convergence of Quantitative and Qualitative Findings

Core result	Quantitative evidence	Qualitative evidence	Integrated meaning
STT supports sound-level pronunciation improvement	Largest difference in gain for segmental accuracy (1.37)	Learners reported noticing wrong words on the screen	Visual transcription appears to promote sound-level self-correction.
STT contributes to clearer overall speech	Overall intelligibility gain difference of 1.05	Learners reported greater confidence and repeated practice	Improved specific pronunciation control may strengthen perceived intelligibility.
STT offers some support for word stress	Word stress gain difference of 0.95	Learners noticed that certain words were not recognized correctly	STT may direct attention toward stress-related pronunciation issues.
STT is less effective for prosody	Small gain differences for rhythm (0.17) and intonation (0.07)	Learners reported that poor rhythm could still be transcribed correctly	STT should not be treated as a complete tool for suprasegmental training.

Note. This table integrates both strands of evidence to strengthen the explanation of the main findings.

The convergence shown in Table 6 enhances the validity of the findings. When the results from the quantitative analysis demonstrated advantages, the learners' narratives provided a logical account for their origins. When the quantitative analysis results were weak, the qualitative results highlighted a relevant drawback. The congruence between the different types of evidence makes it more convincing that the STT technology was most useful for pronunciation assessment in selected aspects.

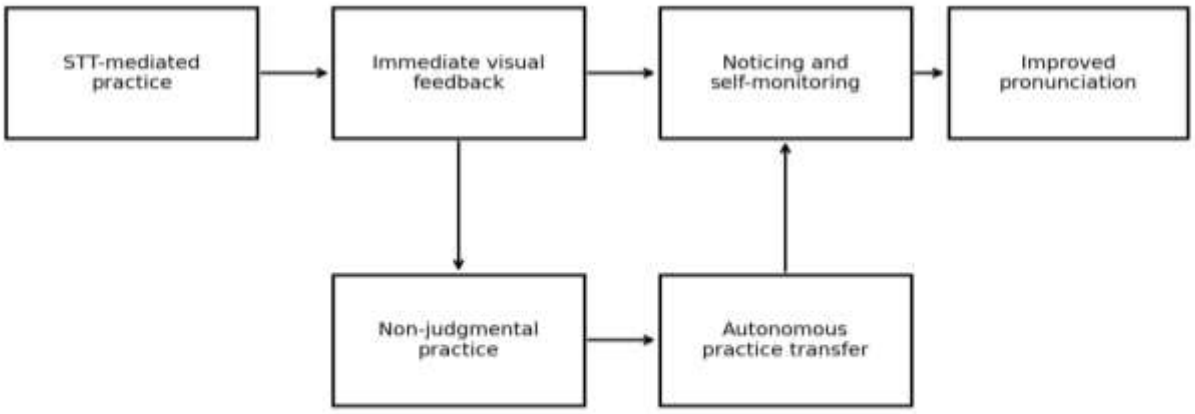

Figure 2. Evidence pathway linking STT feedback with pronunciation development.

Figure 2 summarizes the pathway suggested by the combined results. STT-mediated

Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 4 No. 5 (May) (2026)

practice provides immediate visual feedback, which encourages noticing and self-monitoring. The non-judgmental nature of the tool further promotes repeated practice, while some learners extend this process into autonomous practice outside the classroom. Together, these mechanisms help explain why STT-supported learners demonstrated stronger improvement in selected pronunciation outcomes.

Direct Linkage of Findings with Research Objectives and Research Questions

Study focus	Relevant finding	Support from results	Conclusion
Objective 1 / RQ1: Effectiveness of STT	Experimental group showed greater pronunciation improvement than the control group	Higher gain scores in segmental accuracy, word stress and intelligibility	STT-mediated instruction was more effective for key pronunciation outcomes.
Objective 2 / RQ2: Most and least responsive dimensions	Benefits were strongest for segmental accuracy and intelligibility; weakest for rhythm and intonation	Gain-score and effect-size patterns showed uneven improvement	STT effectiveness is dimension-specific rather than uniform.
Objective 3 / RQ3: Learner perceptions	Learners valued feedback immediacy, non-judgmental practice and self-monitoring	Themes were frequently reported across interviews and journals	STT was perceived as useful and motivating, but still requiring pedagogical support.

Note. The table demonstrates that the analysis directly addresses the objectives and research questions of the study.

Table 7 shows that the findings respond directly to the central concerns of the study. The results establish that STT technology can improve pronunciation outcomes among ESL learners, identify which pronunciation dimensions benefit most and clarify how learners perceive the tool during pronunciation practice. This linkage makes the analysis section more explicitly aligned with the title, objectives and research questions, while also preparing the reader for the subsequent discussion of challenges and frustrations.

Despite these positive patterns, the findings also reveal several limitations in the learner experience, particularly in relation to interpreting system errors, detecting prosodic weaknesses and managing technical constraints. These concerns are discussed in the following subsection.

Challenges and Frustrations

The participants expressed positive feedback but they also pointed to some genuine challenges. The most frequent complaint was the inability to handle prosodic features in a consistent manner. A few learners remarked that they could type a sentence with an incorrect rhythm Andr intonation and the sentence would still be correctly transcribed, which they found confusing because it seemed to imply that their pronunciation was correct when it was not. This is a direct consequence of the quantitative technical limitation of ASR systems: They can cope with prosodically non-target-like utterances without complaint.

Dialogue Social Science Review (DSSR)
www.thedssr.com

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 4 No. 5 (May) (2026)

A second difficulty was that the recognition errors needed interpreting. The visual feedback was generally useful for participants but some were not able to identify which phonological feature led to a misrecognition. Sometimes it was difficult to determine which sound was the issue, especially when the system wrote a different word. This is in line with the theoretical prediction of the noticing hypothesis that noticing an error is a necessary but not sufficient condition: the learning must also detect what must be changed. The message is that the training in the identification of recognition errors is complementary to the training in STT.

Self-Directed Practice Behaviors

An intriguing discovery was the off-schedule usage of the app. Some were willing to extend their practice voluntarily, to practice the pronunciation of words from other classes or to practice academic presentations. This transfer was spontaneous, indicating that the technology had become a self-regulated resource for the learners, in keeping with the sociocultural theory's appropriation concept. It is somewhat like talking to a phone, one user said, when I don't know what to say, I talk to my phone and find out if it does. It has turned into a talking dictionary.

Not all of them were as autonomous. Others used only during mandated sessions and indicated limited independent use. Thematic analysis revealed that the most useful aspect of the app for autonomous learning was the provision of an explicit learning strategy for interpreting errors in recognition tasks and that there was a belief in self-correction. When not supervised, people who were unsure about using the feedback tended to use the technology only when needed. The variant emphasizes the need for training and scaffolding to begin with and throughout.

Discussion

This study investigated the effectiveness of Speech to Text (STT) technology in the area of pronunciation accuracy for adult ESL learners focusing on segmental accuracy, suprasegmental features, intelligibility and the learners' perceptions of automated feedback. In general, the results indicate that STT is suitable as an additional pronunciation-learning tool. The experimental group that practiced with STT-mediated feedback showed significant improvement on a number of significant pronunciation measures compared to the control group. The results are in line with the recent scholarship that confirms that proper and systematic use of automatic speech recognition and computer-assisted pronunciation training can have a positive effect on L2 pronunciation outcomes (Amrate & Tsai, 2025; Ngo et al., 2024).

The largest gain was in segmental accuracy for the experimental group (+1.89 points) compared to the control group (+0.52 points) with a large effect size partial eta squared = .24. The result is significant as it directly reinforces the study's core finding: that STT technology can aid learners' awareness and correction of sound-level pronunciation errors. Misrecognized words are visually represented to the learner as evidence that may show that what he or she produced is not the target. It further builds on previous research that found that ASR feedback could be used to capture the learners' attention to specific phonological contrasts (Ebadi et al., 2019; McCrocklin, 2016).

The findings of word stress also indicate that there is a significant effect in the STT group. The difference between the two groups showed a moderate effect size (partial eta squared = .14), with the experimental group scoring 1.43 points and the control group scoring 0.48 points. The results align with the recent studies that highlight the benefits of using ASR in practice for strengthening learners' self-monitoring skills

ISSN Online: 3007-3154
ISSN Print: 3007-3146

and prompting them to take more careful actions in their autonomous pronunciation tasks (Inceoglu et al., 2024; Leis, 2025).

A key result is overall intelligibility with a significant difference between the experimental and control groups (partial eta squared = .18; an experimental group gain of 1.57 points versus a control group gain of 0.52). This makes a case for communicatively meaningful pronunciation development with STT technology. The results also highlight some of the limitations of STT technology. The differences between the experimental and control groups were much smaller for rhythm and intonation, with effect sizes of only .03 and .02, respectively. The results suggest that STT did not perform particularly well with the more subtle prosodic aspects of speech. The outcome shows an important limitation of the speech-recognition systems: they are optimized mainly for word recognition, but not for the whole spectrum of pronunciation features that lead to expressive fluent speech. The same concern has been expressed in the general CAPT literature (Levis & Sonsaat, 2020; Ngo et al., 2024), which warns that automated feedback should not be seen as a replacement for expert pronunciation feedback.

The qualitative findings add important explanatory depth to the quantitative results. The immediate feedback and the non-judgmental nature of STT-supported practice was most often appreciated by learners. 14/16 participants interviewed described being able to see errors on the screen immediately as useful; and eleven participants described being able to practice without shame or fear of peers correcting them as useful. The perceptions presented here imply that STT technology might be able to lower affective obstacles to pronunciation practice and make more willing to try, repeat and self-correct. The interpretation is consistent with the sociocultural theory that considers the tools to be mediational resources which transform the way learners interact with tasks and manage their learning (Inceoglu et al., 2024; Amrate & Tsai, 2025).

Conclusion

This study compared STT technology with a pedagogical approach as a means to enhance pronunciation among adult ESL learners, which is lacking in empirical CAPT literature. The results indicate that STT technology yields significant segmental accuracy and word stress improvement, which results in better intelligibility without significant improvement in rhythm and intonation. Participants appreciated the responsiveness and non-evaluative nature of automated feedback but they also faced difficulties in understanding the feedback on recognition errors as well as frustration at the lack of sensitivity towards prosodic features of the feedback.

For the teacher, the message is that STT technology can be very useful as an auxiliary device if applied to dimensions where it is reliable such as segmental accuracy and word stress. It should not be used to explain proso-dic features in full and should be continued to be taught in other ways by the teachers. Integrating effective STT into the materials demands beyond just introducing the technology; learners must be trained to understand feedback, identify errors and self-organize practice. The study provides evidence that program administrators can expect tangible benefits resulting from the implementation of STT technology as long as they keep an eye on the pedagogical goals rather than just the innovation hype.

The study additionally has theoretical value. The cross-over with the prediction from the sociocultural theory and the noticing hypothesis reinforces the importance of considering these theories when thinking about how digital tools can facilitate pronunciation development. The finding that those learners who learned to construct effective diagnostic strategies were more likely to learn to use them autonomously

ISSN Online: 3007-3154
ISSN Print: 3007-3146

highlights the sociocultural aspect of the role of tools: they are not effective tools in themselves but effective tools through the processes by which learners appropriate the tools.

There are a number of cautionary statements that should be noted. Due to intact classes, the quasi-experimental design is used which restricts causal inferences. The groups did not differ with regard to any observed variables but unobserved differences in motivation, previous experience with technology use or learning style may have affected the results. The 12-week time frame is enough to see gains but may not reflect the entire process of longer-term STT practice. Few conclusions can be drawn beyond the context of Riphah International University, Islamabad and the population of students in the intensive English program.

Using only one STT application may not apply to other systems that use different algorithms and/or interfaces. The Google Speech-to-Text system is one of the many different systems currently in development and future research should look at whether similar results occur with other systems. All tasks of assessing spoken language through read-aloud and spontaneous speech contribute to important elements of spoken language but do not adequately represent communicative complexity in real settings.

The findings and the limitations suggest directions for future research. Random assignment to create true experimental designs could enhance the conclusions of causation. A study of the longevity of the gains over longer time periods would determine whether the improvement is a lasting learning. Different types of learner groups (younger learners, a range of proficiency levels and different L1 backgrounds) would shed light on boundaries of effectiveness. It would be useful to compare various STT applications and look at interface design characteristics that facilitate learning.

It would also be useful to research scaffolding in teacher-mediated instruction in STT. This study gave teachers initial training and some continuous support but the extent and type of scaffolding is not known. Comparisons of varying the levels of instructional support could hasten the discovery of conditions under which the STT technology is most effective for learners to appropriate. Eye tracking and/or think aloud studies might shed light on the cognitive processes that inform the interpretation of recognition feedback.

It can be concluded that speech to text technology, when used appropriately, can be a valuable tool in supporting the phonological development of ESL learners. It is not a replacement for expert teaching and it will not help with all aspects of pronunciation. As an additional resource for instant visual feedback, helping to alleviate practice anxiety and enabling self-monitoring without intervention, this helps overcome some of the ongoing difficulties in traditional pronunciation contexts. The challenge for educators and researchers is to continue the refinement process of understanding their use, for whom, when and for what learning goals.

References

Amrate, M., & Tsai, P.-H. (2025). Computer-assisted pronunciation training: A systematic review. *ReCALL, 37*(1), 22–42. https://doi.org/10.1017/S0958344024000181

Baker, A. A. (2014). Exploring teachers' knowledge of second language pronunciation techniques: Teacher cognitions, observed classroom practices and student perceptions. *TESOL Quarterly, 48*(1), 136–163. https://doi.org/10.1002/tesq.99

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), 77–101.

ISSN Online: 3007-3154
ISSN Print: 3007-3146

https://doi.org/10.1191/1478088706qp063oa

Celce-Murcia, M., Brinton, D. M., & Goodwin, J. M. (2010). *Teaching pronunciation: A course book and reference guide* (2nd ed.). Cambridge University Press.

Chapelle, C. A. (2001). *Computer applications in second language acquisition: Foundations for teaching, testing and research*. Cambridge University Press.

Chen, N. F. (2016). A study on the effects of computer-assisted pronunciation training on Mandarin learners of English. *Computer Assisted Language Learning*, *29*(5), 1031–1047. https://doi.org/10.1080/09588221.2015.1069361

Derwing, T. M., & Munro, M. J. (2015). *Pronunciation fundamentals: Evidence-based perspectives for L2 teaching and research*. John Benjamins.

Ebadi, E., Naebzadeh, S., & Shahini, G. (2019). The impact of Google Speech-to-Text on the development of EFL learners' pronunciation. *Teaching English with Technology*, *19*(3), 3–18.

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

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A. R., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., & Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal Processing Magazine*, *29*(6), 82–97. https://doi.org/10.1109/MSP.2012.2205597

Inceoglu, S., Chen, W.-H., & Lim, H. (2024). Monitoring student behavior in autonomous automatic speech recognition-based pronunciation practice. *System, 124*, Article 103387. https://doi.org/10.1016/j.system.2024.103387

Isaacs, T. (2018). Shifting tides: The ongoing transformation of pronunciation assessment. *Language Assessment Quarterly*, *15*(3), 237–253. https://doi.org/10.1080/15434303.2018.1486395

Kang, O., Rubin, D., & Pickering, L. (2010). Suprasegmental measures of accentedness and judgments of language learner proficiency in oral English. *Modern Language Journal*, *94*(4), 554–566. https://doi.org/10.1111/j.1540-4781.2010.01091.x

Lantolf, J. P., & Thorne, S. L. (2006). *Sociocultural theory and the genesis of second language development*. Oxford University Press.

Lao-un, J., & Khampusaen, D. (2025). Developing an AI-powered pronunciation application to improve English pronunciation of Thai ESP learners. *Languages, 10*(11), Article 273. https://doi.org/10.3390/languages10110273

Leis, A. (2025). How speech-to-text technology affects pronunciation gains and self-confidence in EFL learners. *Computer Assisted Language Learning*, 1–24. https://doi.org/10.1080/09588221.2025.2534498

Levis, J., & Sonsaat, S. (2020). Pronunciation and technology. In T. Derwing, M. Munro, & R. Thomson (Eds.), *Routledge handbook of contemporary English pronunciation* (pp. 502–520). Routledge.

Liakin, D., Cardoso, W., & Liakina, N. (2015). Learning L2 pronunciation with a mobile speech recognizer: French /y/. *CALICO Journal*, *32*(1), 1–25. https://doi.org/10.1558/cj.v32i1.25900

Liu, Y., Ab Rahman, F., & Mohamad Zain, F. (2025). A systematic literature review of research on automatic speech recognition in EFL pronunciation. *Cogent Education*, *12*(1), Article 2466288. https://doi.org/10.1080/2331186X.2025.2466288

Lyster, R., & Saito, K. (2010). Oral feedback in classroom SLA: A meta-analysis. *Studies in Second Language Acquisition*, *32*(2), 265–302.

ISSN Online: 3007-3154
ISSN Print: 3007-3146

https://doi.org/10.1017/S0272263109990520

McCrocklin, S. M. (2016). Pronunciation learner autonomy: The potential of automatic speech recognition. *TESOL Journal*, *7*(3), 668–685. https://doi.org/10.1002/tesj.246

Neri, A., Mich, O., Gerosa, M., & Giuliani, D. (2008). The effectiveness of computer-assisted pronunciation training for foreign language learning by children. *Computer Assisted Language Learning*, *21*(5), 393–408. https://doi.org/10.1080/09588220802447871

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

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

Schmidt, R. (2001). Attention. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 3–32). Cambridge University Press.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.