

Automated Speech Recognition Technology for Improving English Pronunciation Skills

Shazia Khanum Yunus^{1*} and Neha Dey²

^{1*}English Language Trainer, Department of Common Core Curriculum, CMR University, Bangalore, Karnataka, India. E-mail: shaziya.lsi@cmr.edu.in

²Career Counsellor and Soft Skills Trainer, Department of Common Core Curriculum, CMR University, Bangalore, Karnataka, India. E-mail: neha.d@cmr.edu.in

Abstract: Pronunciation continues to be a problem for EFL learners since conventional classrooms are unable to provide immediate and personalized feedback in these cases. This research explores the application of Automated Speech Recognition (ASR) technology as an additional tool in pronunciation teaching, based on Second Language Acquisition (SLA) theories, which include Noticing, Output, and Learner Autonomy Hypotheses. Using a descriptive quantitative approach, 30 EFL undergraduate learners at an intermediate level (N=30) underwent a 6-week intervention, where practiced using ASR guidance (about 20 min. 3 times a week). The performance of learners before and after the intervention was assessed based on a 100-point rating system in five aspects of pronunciation: segmental correctness, word stress, sentence intonation, speech fluency, and speaker confidence. The results of data analysis indicate uniform improvement of mean scores in all analyzed metrics. The greatest increase was found in the metric of self-reported confidence of speakers (+24 points, from 54 to 78), thus indicating the affective value of low-pressure automated assessment. The metrics of word stress (55 to 74) and sentence intonation (52 to 71), along with speech fluency (60 to 79), have increased in unison by +19 points. The segmental accuracy score gained by +18 points (58 to 76). ASR tools are effective in fostering independent phonological development and confidence in communication. Although ASR technology is unable to fully substitute for human-based instruction because of the difficulties with assessment of suprasegmental components, it is an extremely valuable supplementary tool in language education curriculum.

Key Words: automated speech recognition, english pronunciation, computer-assisted language learning, second language acquisition, learner autonomy, phonological accuracy, EFL pedagogy

(Received: 19 June 2024; Revised: 31 July 2024; Accepted: 04 September 2024; Published: 30 September 2024)

Introduction

Pronunciation continues to be one of the most resistant aspects of learning English as a second or foreign language, often remaining an issue even after learners acquire a competent level of grammar and vocabulary. Unlike other areas of language skills, pronunciation is closely linked to the phonological patterns of language acquisition at the time of the first language learning process, and adult learners tend to face difficulty both recognizing and fixing their pronunciation errors. Study notes that the combination of such factors as the persistent problem of pronunciation and its infrequent practice in the main EFL classroom leads to the necessity of developing additional means of improving this skill (Sun, 2023). Automated Speech Recognition (ASR) technology has become one of the most promising solutions to this problem.

Explanation of Automated Speech Recognition (ASR) Technology

At its core, ASR technology means computerized systems that have the ability to receive speech as an input, transform the speech into a digital format, and compare it to the model of pronunciation of the target language. The application of ASR technology can be illustrated by the example of ELSA Speak, which provides feedback at the phoneme level and word level in real time (Kholis, 2021). This type of feedback is instantaneous and non-threatening because it is not provided by a human being. That is why ASR technology helps learners to practice their speaking skills and correct mistakes independently at their own pace, and this characteristic of ASR technology is considered one of the main educational benefits of ASR technology over traditional classroom correction (Elimat & AbuSeileek, 2014).

Importance of English Pronunciation Skills

The importance of Pronunciation is much more than just accurate production of sounds. Effective pronunciation is crucial for effective oral communication, forms the basis of listeners' perception of a speaker's proficiency, and plays a critical role in developing a learner's confidence in the practical use of the language. According to the research, lack of sufficient pronunciation training and resulting anxiety often led to complete reluctance of learners to practice spontaneously, thus delaying language acquisition as a whole (Evers & Chen, 2022). Pronunciation, therefore, serves as a technical skill key to effective communication, confidence, and fluency.

How ASR Technology can Improve Pronunciation Skills

This research will examine how far ASR technology may prove to be an efficient tool for helping English language learners develop better pronunciation skills through its application as an auxiliary aid. The study of the use of ASR-based software by Iraqi EFL learners shows that consistent practice with the use of this tool can yield good results with regard to developing learner pronunciation within a relatively short period of time, thereby providing an empirical foundation for the educational claims made regarding the technology (Sidgi & Shaari, 2017).

Here is a summary of the paper's unique contributions:

- Synthesizes Second Language Acquisition (SLA) theories, specifically the Input, Noticing, Output, and Learner Autonomy hypotheses, to provide a solid theoretical foundation for *why* and *how* ASR technology enhances pronunciation practice.
- Demonstrates through a controlled 6-week study ($N = 30$) that structured ASR practice yields measurable quantitative gains across both segmental (individual sounds) and suprasegmental (word stress, sentence intonation, fluency) dimensions, alongside self-rated learner confidence.
- Establishes a nuanced position on ASR in EFL pedagogy, defining it not as a replacement for human instruction, but as an effective, low-anxiety supplementary tool for autonomous practice.

Following on from the above, the structure of this paper. after the introduction in part 1, Section 2 presents a review of the relevant literature, putting the current study into context; Section 3 outlines the methodology of the current study; Section 4 presents the results of the ASR technology and Section 5 discusses interpret the findings; and the paper closes with a conclusion in Section 6.

Literature Review

Studies on the Effectiveness of ASR Technology for Language Learning

A substantial number of research papers have explored the efficacy of ASR technology in general and specifically for the purpose of pronunciation training in language learning. presents a comprehensive theory on how ASR pronunciation training functions, putting it in the context of the larger field of computer-assisted pronunciation training and suggesting that the key advantage of ASR pronunciation training is its ability to offer a large amount of feedback instantly, which cannot be done by human teachers because of constraints on time and the number of students in the class (Cucchiarini & Strik, 2017). The possibility of giving instant feedback has been verified in further empirical studies (Kim, 2023). A comparative analysis of automated and human rater evaluations of pronunciations has revealed that although ASR evaluation may deviate from human judgment in evaluating the suprasegmental aspects of speech, including intonation, it still may come close enough to human evaluations to be pedagogically effective, even in the context of gross segmental mistakes, even though some differences may be seen at the level of suprasegmental aspects like intonation. This issue is further discussed in the study, in which the link between ASR practice and learner autonomy has been investigated, and it has been found that learners performing self-directed ASR practice feel more ownership in their pronunciation practice and are more ready to speak out of the class environment (McCrocklin, 2016). A similar study also highlights the improvements in learners' pronunciation accuracy after ASR practice, and the primary value of ASR technology is considered to be in its ability to provide learners with a large amount of practice that it makes possible (Liu et al., 2019).

Comparison of Different ASR Tools Available for English Pronunciation Practice

In addition to establishing overall effectiveness, there has also been research comparing the impact of varying ASR-based technologies and delivery methods on learning results. For instance, with regard to mobile-assisted pronunciation learning, it was found that the combined use of ASR feedback and peer feedback yields better learning outcomes than either of the two sources of feedback separately, indicating that ASR technology is more effective when used alongside, rather than instead of, social feedback (Dai & Wu, 2023). In a similar vein, in another study, it was found that not all learners benefit equally from ASR-based pronunciation training software, as learners with various learning styles react differently to the rhythm, delivery, and feedback type used by the technology (Evers & Chen, 2021). Analyzing the user experience with a mobile app for pronunciation practice, which was developed using ASR technology, it was similarly revealed that user satisfaction and the usefulness of the technology are influenced by learners' level of familiarity with mobile technology and tolerance for ASR-related misrecognitions (Ahn & Lee, 2016). From these studies, it can be concluded that although all ASR-based tools operate based on the same mechanism, their pedagogical efficacy differs significantly depending on their design, delivery, and integration into learners' practice routine.

Challenges and Limitations of Using ASR Technology for Pronunciation Improvement

Nevertheless, there are also significant problems and limitations in the application of ASR technology for pronunciation training. While reviewing an ASR-based website designed for oral skills training, mentions that the accuracy rate may decrease dramatically when learners use heavily accented speech or have atypical intonation patterns, which in turn results in the provision of inaccurate or even demotivating feedback (Chen, 2011). In their study of a controlled computer-assisted pronunciation instrument, report that improvements made by learners through the use of the system are relatively small and vary from person

to person; moreover, learners need to use the instrument regularly to achieve any noticeable results (Tejedor-García et al., 2020). Analyzing ASR-based pronunciation assessment systems that are based on deep learning techniques, points out that despite all technological advances, these systems still do not manage to take into account suprasegmental aspects of pronunciation, which actually have the greatest influence on intelligibility (Xu, 2022). This implies that although ASR technology has great potential from the pedagogical perspective, it still needs to be seen as an element of a larger, human-supported teaching process for pronunciation instruction rather than as a self-sufficient solution.

Previous literature notes that ASR technology increases learners' autonomy and enables instant, large amounts of feedback; however, the technologies sometimes have difficulty dealing with suprasegmental aspects such as rhythm and stress. The results obtained from the study provide confirmation to previous literature showing that there was a considerable improvement in segmental accuracy, word stress, intonation, and confidence after practicing ASR. Finally, both previous literature and research results confirm the use of ASR as a supplementary tool and not a replacement for human instruction.

Methodology

ASR within Language Acquisition

ASR-based pronunciation training within the framework of existing theories of second language acquisition may facilitate understanding of its pedagogical value. Theories of language acquisition based on the role of input postulate that learners need a great deal of comprehensible input in order to be able to learn a target language productively. The role of ASR-based tools in this process lies in exposing learners to comprehensible input through model pronunciation of target words and phrases, thus making it possible for learners to compare their output to the phonological model.

It is also important to mention the idea that learners should consciously notice the discrepancy between their output and the target model in order for this discrepancy to be eliminated. Correction in a traditional classroom environment is usually sporadic and easily lost in the course of the class, while ASR feedback singles out particular sounds, words, or stresses and makes them available for analysis by learners. The possibility of making pronunciation errors visible instead of just audible is probably the most crucial function of ASR technology that supports learning.

Another relevant line of theorization is concerned with the importance of producing, as opposed to just comprehending, language. Pronunciation is inevitably a productive skill, and theories focused on the importance of production emphasize that proficiency at speaking emerges from the repeated effort, errors, and corrective attempts that come from speaking. ASR software provides learners with the means of practicing their speaking ability without any stress related to addressing a classmate or instructor since have to repeat speaking into a microphone and get immediate feedback.

Finally, theories of learner autonomy suggest that language skills develop best when learners actively participate in the process of their acquisition. Since ASR software can be applied in the learners' personal time, independently of any classroom environment make such participation possible. In other words, theories of learner autonomy suggest that ASR technology enables learners to take an active part in the process of acquiring pronunciation skills. Thus, may conclude that this type of software is consistent with several major principles of language pedagogy, which explains its potential to enhance pronunciation learning.

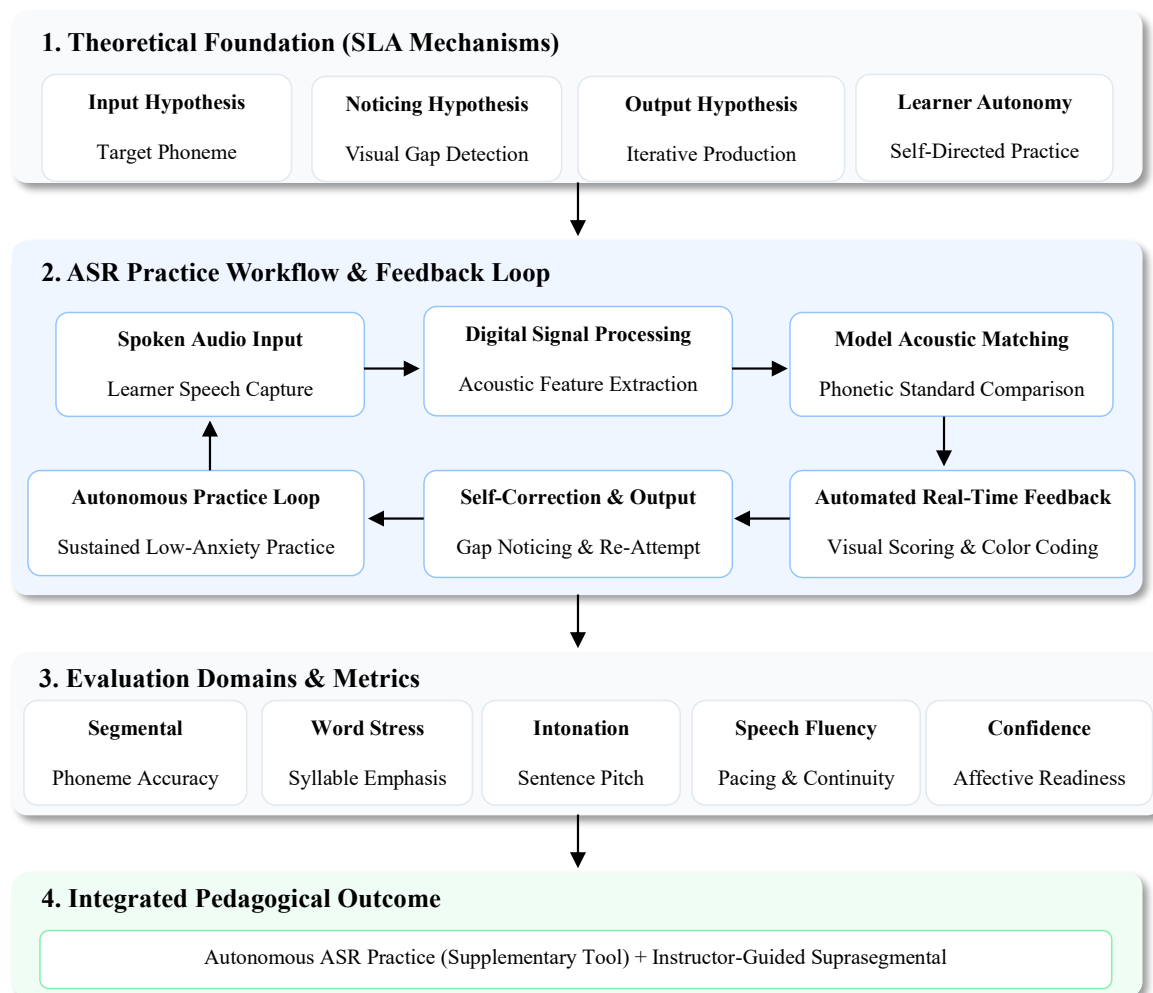


Figure 1: Semantic-aware Edge Intelligence Pipeline for Adaptive QoE Optimization

Figure 1 shows the relationship between the application of SLA theories and automated speech processing that helps in the pronunciation improvement process. The diagram clearly reflects how theoretical assumptions like the production of output and noticing the gaps are translated into an iterative and anxiety-free ASR process. In doing so, it clearly illustrates how automated speech recognition operates in practice as a tool that works side by side with the traditional role of an instructor's guidance.

Participants Involved in the Study

The present study is based on a small-scale descriptive quantitative research design and the teaching effect of ASR pronunciation practice. The participants of the study include thirty intermediate-level learners of English as a foreign language from an undergraduate English language program and aged between eighteen and twenty-two years. These participants were chosen from a similar educational background, and all had undergone similar formal instruction in English without ever practicing with ASR pronunciation software in a sustained manner.

ASR Tool Used for the Research

The ASR tool selected for the experiment is an existing mobile application for pronunciation training, which provides a score of learner speech precision at the phoneme, word, and short sentence level, giving a numerical accuracy score, along with the colored representation of mispronunciations. The tool was chosen due to its availability, the ability to work independently from the classroom setting, and the

possibility to provide stable scores throughout the experiment for comparison purposes between learners' performance at different points in the study.

Procedure for Assessing English Pronunciation Skills Before and After Using ASR Technology

The process was divided into three main phases. The first phase involved carrying out a pre-test, during which the participants were asked to read a standard passage and a collection of target sentences that could be used to generate different English consonant and vowel sounds, word stress, and intonation patterns, which were evaluated on five main parameters identified in this study. In the second phase, participants took part in guided ASR-based pronunciation training sessions lasting about twenty minutes per session, three times a week for six weeks, using the pronunciation activities related to the sounds and patterns tested in the pre-test. Finally, in the third phase, the participants took a post-test using a similar passage and target sentences evaluated based on the five parameters, allowing a direct comparison of pronunciation performance before and after the practice period.

Results

Analysis of the Data Collected from the Participants

The results of the analysis demonstrate positive improvements in all five categories after ASR training for six weeks. The biggest improvement is shown by speaker confidence, which increased by +24 points (54 to 78). This result proves the importance of psychologically secure feedback for learning. Suprasegmental categories of word stress and sentence intonation improved identically by +19 points. Segmental accuracy demonstrated the smallest increase, but still +18 points (58 to 76), since the learners' initial levels were high. Thus, it can be stated that automated feedback helps to improve sound accuracy and speech melody.

Comparison of Pronunciation Skills Before and After Using ASR Technology

After the six weeks' intervention, there were performance improvements in all areas after the post-test assessment. The most improved area was the confidence level of the speakers, which went from 54 to 78 (+24). Supra-segmental areas of word stress (+19), sentence intonation (+19), speech fluency (+19), and segmental accuracy (+18) all made notable improvements, confirming ASR's efficacy for phonological development.

Table 1: Mean Pronunciation Scores Before and After ASR-based Practice (N = 30)

Pronunciation Parameter	Pre-Test Mean Score (/100)	Post-Test Mean Score (/100)	Mean Gain
Segmental Accuracy	58	76	+18
Word Stress	55	74	+19
Sentence Intonation	52	71	+19
Speech Fluency	60	79	+19
Speaker Confidence (self-rated)	54	78	+24

Table 1 is the representation of the mean scores of the participants' performances before and after six weeks of the ASR training program. Improvements were achieved by all five elements evaluated in this study, and the largest improvements were made in the areas of confidence and mastery of sentence intonation.

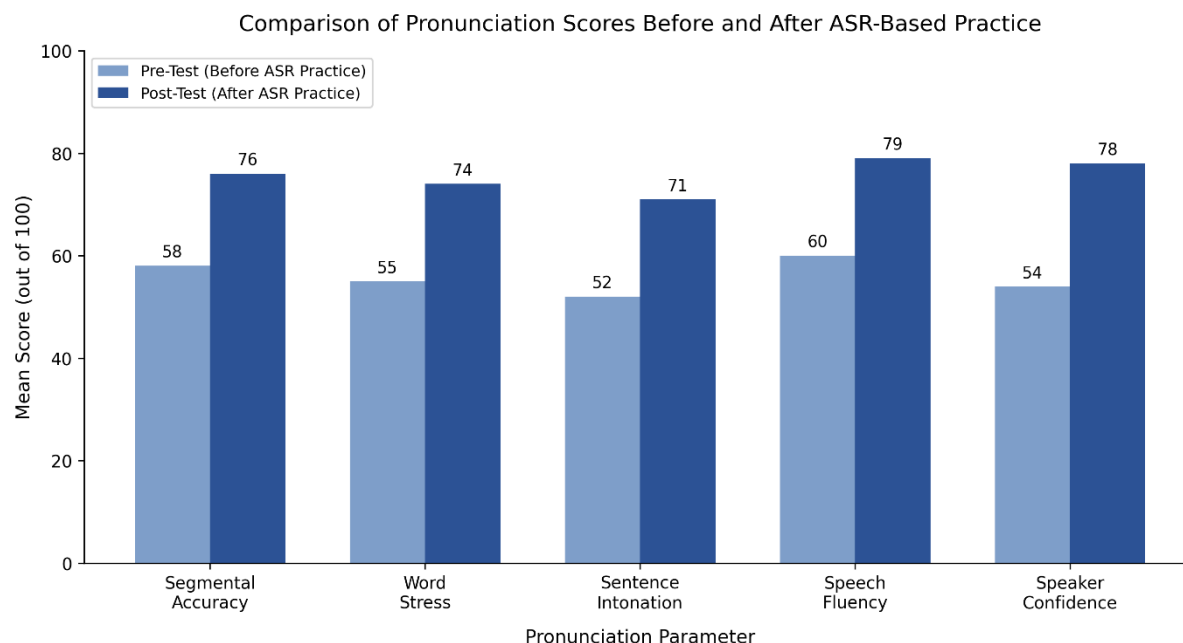


Figure 2: Comparison of Pronunciation Scores Before and After ASR-based Practice

Figure 2 shows the same results graphically, showing the consistently increasing mean results across all five criteria after the practice phase.

The smallest gain observed in absolute terms was observed in the segmental accuracy criterion – the one in which the participants initially demonstrated relatively high results in the pre-test. At the same time, the greatest gain observed was in the speaker confidence criterion – self-assessed rather than externally evaluated. This suggests that the positive impact of regular practice in a stress-free environment can be considered no less important than the phonological gains. The gains in the word stress and sentence intonation criteria – the two suprasegmental criteria – were of similar magnitude and greater than those for segmental accuracy. This means that ASR practice was equally effective for rhythm and melody as for segmental accuracy, meaning that ASR training proved to be as helpful in mastering speech rhythm and melody as in pronouncing individual sounds correctly.

Discussion

These results from all five factors have many implications not only for English language learners but also for those who teach them. For learners, these results imply that consistent and independent practice through an ASR system may result in noticeable improvements in pronunciation in a rather short period of time without constant monitoring by teachers. This is especially important for learners whose classes provide little time to focus on pronunciation skills development, or who lack the opportunity to practice speaking with a native speaker of English.

For educators, the findings point toward the value of incorporating ASR-based tools as a structured, supplementary component of English language instruction rather than as an occasional or optional activity. Given that participants in this study achieved measurable gains through relatively brief, regular practice sessions, instructors might consider assigning short ASR-based exercises as homework tasks that complement, rather than replace, in-class pronunciation instruction. Equally, given that the literature reviewed earlier in this paper cautions that ASR tools can struggle with learners' regional accents and with the finer suprasegmental features of speech, instructors should continue to provide periodic human

feedback alongside automated practice, ensuring that any recurring or systemic misjudgments produced by the software are identified and corrected are identified and corrected.

The future direction for research in this topic could include various extensions to the current experiment: conducting an experiment over a longer timeframe to see if the gains made through the practice are sustained or peak at some point; evaluating the comparative performance of various ASR software; and researching the role played by the learners' attitude towards technology, along with familiarity with mobile apps. These areas of investigation could lead to better recommendations for practitioners trying to incorporate ASR technology in their classrooms.

Conclusion

This study provides evidence of the effective use of Automated Speech Recognition (ASR) technology as a useful tool for English pronunciation instruction. After the completion of a carefully planned six-week ASR program targeting intermediate learners of English (N=30), empirical evaluation showed significant improvement in all five parameters investigated. There was a remarkable rise of 24 points (+24) in the self-assessment scores related to the speaker's confidence level, indicating the immense affective advantage of low-anxiety and automated practice conditions. There were equal rises of 19 points (+19) in the area of suprasegmentals and rhythm, which include word stress (55 to 74), sentence intonation (52 to 71), and speech fluency (60 to 79). This study proves that real-time, visual, and non-evaluative ASR provides a great opportunity to implement the principles of Second Language Acquisition such as gap noticing, output structure repetition, and learner autonomy. Nonetheless, since the technology is known to have its limitations in terms of dealing with thick accents and suprasegmentals, ASR should be regarded as an auxiliary, learner-centered tool that does not substitute for human interaction entirely. Future research will build on this preliminary data through the undertaking of longitudinal studies that aim to uncover the retention rate of phonological improvements after six weeks. Further research into different ASR algorithms, like the comparison of deep-learning-based algorithms and conventional acoustic modeling, among others, will help form actionable recommendations on how to incorporate mobile pronunciation tools into communicative language courses curricula.

References

- Ahn, T. Y., & Lee, S. M. (2016). User experience of a mobile speaking application with automatic speech recognition for EFL learning. *British Journal of Educational Technology*, 47(4), 778-786. <https://doi.org/10.1111/bjet.12354>
- Chen, H. H. J. (2011). Developing and evaluating an oral skills training website supported by automatic speech recognition technology. *ReCALL*, 23(1), 59-78. <https://doi.org/10.1017/S0958344010000285>
- Cucchiarini, C., & Strik, H. (2017). Automatic speech recognition for second language pronunciation training. In *The Routledge handbook of contemporary English pronunciation* (pp. 556-569). Routledge. <https://doi.org/10.4324/9781315145006>
- Dai, Y., & Wu, Z. (2023). Mobile-assisted pronunciation learning with feedback from peers and/or automatic speech recognition: A mixed-methods study. *Computer Assisted Language Learning*, 36(5-6), 861-884. <https://doi.org/10.1080/09588221.2021.1952272>
- Elimat, A. K., & AbuSeileek, A. F. (2014). Automatic speech recognition technology as an effective means for teaching pronunciation. *The jalt call Journal*, 10(1), 21-47. <https://doi.org/10.29140/jaltcall.v10n1.j166>

- Evers, K., & Chen, S. (2021). Effects of automatic speech recognition software on pronunciation for adults with different learning styles. *Journal of Educational Computing Research*, 59(4), 669-685. <https://doi.org/10.1177/0735633120972011>
- Evers, K., & Chen, S. (2022). Effects of an automatic speech recognition system with peer feedback on pronunciation instruction for adults. *Computer Assisted Language Learning*, 35(8), 1869-1889. <https://doi.org/10.1080/09588221.2020.1839504>
- Kholis, A. (2021). Elsa speak app: automatic speech recognition (ASR) for supplementing English pronunciation skills. *Pedagogy: Journal of English Language Teaching*, 9(1), 1-14. <https://doi.org/10.32332/joelt.v9i1.2723>
- Kim, M. (2023). Digital enhancement of pronunciation assessment: Automated speech recognition and human raters. *Phonetics and Speech Sciences*, 15(2), 13-20. <https://doi.org/10.13064/KSSS.2023.15.2.013>
- Liu, X., Xu, M., Li, M., Han, M., Chen, Z., Mo, Y., ... & Liu, M. (2019). Improving English pronunciation via automatic speech recognition technology. *International Journal of Innovation and Learning*, 25(2), 126-140. <https://doi.org/10.1504/IJIL.2019.097674>
- McCrocklin, S. M. (2016). Pronunciation learner autonomy: The potential of automatic speech recognition. *System*, 57, 25-42. <https://doi.org/10.1016/j.system.2015.12.013>
- Sidgi, L. F. S., & Shaari, A. J. (2017). The usefulness of automatic speech recognition (ASR) eyespeak software in improving Iraqi EFL students' pronunciation. *Advances in language and literary studies*, 8(1), 221-226. <http://doi.org/10.7575/aiac.all.v.8n.1p.221>
- Sun, W. (2023). The impact of automatic speech recognition technology on second language pronunciation and speaking skills of EFL learners: a mixed methods investigation. *Frontiers in psychology*, 14, 1210187. <https://doi.org/10.3389/fpsyg.2023.1210187>
- Tejedor-García, C., Escudero-Mancebo, D., Cámara-Arenas, E., González-Ferreras, C., & Cardeñoso-Payo, V. (2020). Assessing pronunciation improvement in students of English using a controlled computer-assisted pronunciation tool. *IEEE Transactions on Learning Technologies*, 13(2), 269-282. <https://doi.org/10.1109/TLT.2020.2980261>
- Xu, Y. (2022). English speech recognition and evaluation of pronunciation quality using deep learning. *Mobile Information Systems*, 2022(1), 7186375. <https://doi.org/10.1155/2022/7186375>