APPLYING SPEECH RECOGNITION TECHNOLOGY IN TEACHING SPEAKING SKILL AT THE UNIVERSITY OF LABOUR AND SOCIAL AFFAIRS

***Lai Minh Thu*[1]**
Email: minhthulai@gmail.com

Received: 07/03/2025
Revised: 04/09/2025
Accepted: 22/09/2025

DOI: 10.59266/houjs.2025.748

Abstract: *This study investigates the effectiveness of using Automatic Speech Recognition (ASR) technology, specifically the ELSA Speak application, to enhance English pronunciation among non-English major students at the University of Labour and Social Affairs. A quasi-experimental design was employed with 60 first- and second-year students divided into two groups: an experimental group using ELSA Speak for six weeks and a control group receiving traditional instruction. The results revealed that the experimental group showed significantly greater improvements in segmental accuracy, suprasegmental features, and overall intelligibility. Learner feedback also indicated high satisfaction with the app's usefulness, ease of use, and timely feedback. These findings highlight the potential of ASR integration as an effective supplementary tool for pronunciation instruction in Vietnamese public university settings and suggest promising directions for technology-enhanced, personalized, and autonomous language learning.*

Keywords: *English pronunciation, speech recognition, ELSA Speak, educational technology, self-directed learning, English language teaching*

I. Introduction

English has become a critical skill for students in Vietnam, particularly in fields related to labor, human resource management, and social work. As the country integrates more deeply into the global economy, the ability to communicate effectively in English is increasingly demanded in both domestic and international labor markets. Among the four core language skills, pronunciation plays a fundamental role in oral communication. However, it remains one of the most challenging aspects of English learning for Vietnamese students due to phonological differences between Vietnamese and English, as well as limited opportunities for feedback and practice.

At the University of Labour and Social Affairs (ULASA), students are primarily trained for professions that require interpersonal interaction, including social service delivery, labor consulting, and administrative management. Therefore, intelligible and confident English speech is not merely

[1] University of Labour and Social Affairs

an academic target but a vocational necessity. Despite this importance, English pronunciation instruction at the university is often constrained by large class sizes, insufficient contact hours, and a lack of individualized feedback. These constraints call for innovative teaching strategies and technological support to enhance the effectiveness of pronunciation training.

In recent years, speech recognition (SR) technology has emerged as a promising tool in language learning, offering real-time feedback on learners' spoken input. Applications such as ELSA Speak, Google Speech-to-Text, and Microsoft's Azure STT enable students to receive instant evaluations of their pronunciation at the word and phoneme level. Unlike traditional classroom-based pronunciation drills, SR tools promote learner autonomy, self-monitoring, and sustained engagement beyond classroom hours. Nevertheless, the effectiveness of such tools in improving learners' pronunciation accuracy, particularly in Vietnamese university settings, remains underexplored.

This study aims to examine the impact of integrating SR technology into English pronunciation instruction for first- and second-year non-English major students at ULASA. Specifically, it investigates whether the use of SR applications can lead to measurable improvements in students' pronunciation performance and how learners perceive the role of this technology in their learning process. By addressing these questions, the study seeks to contribute empirical evidence to the growing body of research on technology-enhanced language learning and provide pedagogical insights for English instructors in similar educational contexts.

II. Theoretical background

Recent developments in educational technology have significantly reshaped the landscape of English language teaching, particularly in pronunciation instruction. In the past, pronunciation was often relegated to the periphery of language curricula, receiving minimal classroom attention due to time constraints and the dominance of grammar-focused approaches. However, with the increasing recognition of pronunciation as a core component of communicative competence and the advent of mobile-based automatic speech recognition (ASR) tools, new opportunities have emerged for learners to engage in personalized, autonomous, and feedback-rich pronunciation practice.

This section reviews relevant theoretical and empirical foundations underpinning the application of ASR technology in English pronunciation teaching. It is structured around five key dimensions: (1) the role of pronunciation in English language learning, (2) an overview of ASR technology and its pedagogical potential, (3) empirical studies on ASR effectiveness, (4) theoretical frameworks supporting ASR integration, and (5) limitations and considerations in using ASR tools.

2.1. The role of pronunciation in EFL learning

Pronunciation plays a vital role in the intelligibility and effectiveness of spoken communication. Research by Derwing and Munro (2015) has shown that comprehensibility, or how easily a listener can understand speech, is more relevant to real-world communication than achieving native-like pronunciation. For Vietnamese learners of English, persistent pronunciation problems are frequently attributed to negative L1 transfer, lack of phonological awareness, and limited access to corrective feedback in crowded classroom settings (Nguyen & Newton, 2020). These challenges underline the necessity of adopting new methods that provide learners with targeted and frequent pronunciation practice.

2.2. Automatic speech recognition in language learning

2.2.1. Definition and functionality

Automatic Speech Recognition (ASR) refers to computer-based systems that transcribe spoken input into text and can provide analytic feedback on pronunciation. In language learning, ASR tools such as ELSA Speak, Google Speech-to-Text, and Microsoft's Azure API allow learners to practice speaking and receive immediate, individualized feedback on segmental and suprasegmental features of pronunciation. These tools typically highlight errors at the phoneme level, enabling learners to self-correct and track improvement over time (McCrocklin, 2016).

2.2.2. Benefits for learners

ASR technology offers numerous advantages in EFL contexts. It supports learner autonomy, extends learning beyond the classroom, and allows unlimited repetition without judgment. These features are particularly beneficial in Vietnamese higher education, where English instruction is often time-limited and heavily exam-oriented.

2.3. Empirical evidence of ASR effectiveness

Numerous studies have documented the positive impact of ASR on learners' pronunciation skills. Elimat and AbuSeileek (2014) found that Jordanian EFL students using ASR-based systems made significant gains in both segmental accuracy and suprasegmental fluency. In Vietnam, Pham and Pham (2025) conducted a large-scale survey involving 205 students and reported high satisfaction levels with ELSA Speak, citing its ease of use and immediate corrective feedback as the main strengths.

A four-month experimental study by Nguyen et al. (2025) further confirmed the efficacy of ELSA Speak in improving pronunciation accuracy among 37 Vietnamese undergraduates. Other international studies, such as Anggraini (2022) and Sholekhah and Fakhrurriana (2023), echoed similar findings, showing that ASR integration led to statistically significant improvements in learners' articulation, stress patterns, and speech confidence.

2.4. Pedagogical frameworks supporting ASR use

The application of ASR in EFL instruction aligns closely with established models of technology-mediated learning, particularly Computer-Assisted Language Learning (CALL) and Mobile-Assisted Language Learning (MALL). CALL emphasizes the use of computers to facilitate language development through interactive and feedback-driven activities (Chapelle & Jamieson, 2008). MALL extends these principles to mobile platforms, providing learners with access to learning materials anytime and anywhere (Kukulska-Hulme & Viberg, 2018).

Furthermore, the integration of gamified features, adaptive feedback, and progress visualization common in tools like ELSA Speak contributes to increased learner motivation and reduced anxiety. These affordances are especially valuable in pronunciation training, where many learners feel self-conscious about speaking aloud.

III. Research methodology

This study employed a quasi-experimental design to evaluate the effectiveness of automatic speech recognition (ASR) technology in improving English pronunciation among non-English major students at the University of Labour and Social Affairs (ULASA). A total of 60 first- and second-year students from the Faculty of Social Work and Human Resource Management were purposively selected based on their pre-intermediate English proficiency and willingness to participate in technology-enhanced learning activities.

Participants were randomly divided into two groups: the experimental group (n = 30), which received pronunciation practice using the ELSA Speak mobile application, and the control group (n = 30), which received traditional teacher-led pronunciation instruction. The intervention lasted for six weeks, during which students in the experimental group were required to practice with ELSA for at least 20 minutes per day, five days per week. The control group followed the standard curriculum without technological intervention.

Data collection instruments included a pre-test and post-test using a standardized pronunciation rubric assessing segmental features (consonants and vowels), suprasegmental features (stress and intonation), and overall intelligibility. In addition, a short learner perception survey was administered to the experimental group to explore attitudes toward the use of ASR in pronunciation learning.

Quantitative data were analyzed using paired sample and independent sample t-tests to determine within-group improvement and between-group differences. Qualitative responses from the survey were thematically coded to identify emerging patterns related to learner motivation, ease of use, and perceived effectiveness of the ASR tool.

This mixed-methods approach provides a balanced evaluation of both performance outcomes and learner experiences, offering valuable insights into the integration of ASR technology in EFL pronunciation instruction in a Vietnamese university context.

IV. Findings

This section presents the results of the quasi-experimental study evaluating the impact of automatic speech recognition (ASR) technology, specifically the ELSA Speak application, on improving the English pronunciation of students at the University of Labour and Social Affairs. The findings are structured into two parts: (1) quantitative results from pre-test and post-test comparisons and (2) qualitative feedback collected via student surveys on their experiences with the ASR tool.

4.1. Quantitative results

All participants (N = 60) completed a pronunciation test before and after the 6-week intervention period. The assessment rubric evaluated three dimensions: segmental accuracy (consonant and vowel production), suprasegmental features (stress, rhythm, intonation), and overall intelligibility. Each dimension was scored on a scale from 1 (very poor) to 10 (excellent). The descriptive statistics and paired t-test results are summarized below.

Table 1. Pre-test and Post-test results for experimental and control groups

Pronunciation component	Group	Pre-test mean (SD)	Post-test mean (SD)	Mean gain	T-value	P-value
Segmental Accuracy	Experimental	5.86 (0.92)	7.68 (0.85)	+1.82	8.315	< 0.001**
	Control	5.94 (0.88)	6.41 (0.79)	+0.47	2.104	0.043*
Suprasegmental Features	Experimental	5.23 (1.03)	6.91 (1.02)	+1.68	7.689	< 0.001**
	Control	5.31 (0.96)	5.74 (0.91)	+0.43	1.844	0.073
Intelligibility	Experimental	5.71 (0.84)	7.42 (0.87)	+1.71	7.960	< 0.001**
	Control	5.69 (0.89)	6.03 (0.86)	+0.34	1.576	0.126

**p < 0.05; **p < 0.01*

Interpretation of Results

- The experimental group demonstrated statistically significant improvements across all three components, with p-values < 0.001 for segmental accuracy, suprasegmental features, and intelligibility.

- The control group showed only modest gains, and only segmental improvement reached statistical significance (p = 0.043).

- The mean gain in the experimental group for segmental accuracy (+1.82), suprasegmental features (+1.68), and intelligibility (+1.71) was nearly four times greater than the corresponding improvements in the control group.

- These findings support the hypothesis that ASR-based pronunciation training provides more effective and intensive pronunciation enhancement than traditional instruction alone.

4.2. Distribution of individual scores

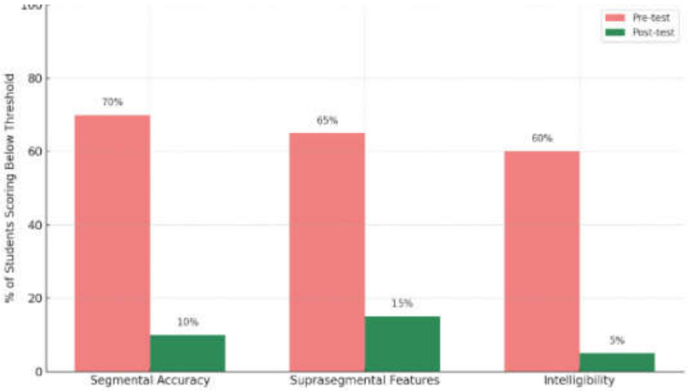

Figure 1. Score distribution (Experimental group)

To further illustrate the improvement, the following chart compares the distribution of pronunciation scores before and after the intervention for the experimental group.

4.3. Qualitative feedback from student survey

To complement the quantitative data, a post-intervention survey was administered to students in the experimental group (n = 30) to gather their reflections on the learning experience. The survey used a 5-point Likert scale (1 = Strongly disagree, 5 = Strongly agree) and included open-ended items.

Table 2. Student perceptions of ASR use (ELSA Speak)

Survey statement	Mean score	SD
The app helped me detect my pronunciation errors more clearly	4.53	0.57
I felt more confident speaking English after using the app	4.38	0.62
The app’s feedback was timely and easy to understand	4.46	0.55
I was more motivated to practice pronunciation using the app	4.29	0.63
I prefer using ELSA Speak to traditional classroom drills	4.12	0.73
I will continue to use the app even after this course ends	4.47	0.59

The post-intervention survey results indicate a high level of student satisfaction and positive attitudes toward the use of ELSA Speak in pronunciation practice. The data suggests that the integration of ASR technology not only improved performance but also enhanced learner motivation, awareness, and confidence.

4.3.1. Increased error awareness and accuracy of feedback

The statement *"The app helped me detect my pronunciation errors more clearly"* received the highest average score (M = 4.53, SD = 0.57). This suggests that students highly valued the immediate and detailed phoneme-level feedback provided by the app, which enabled them to recognize and self-correct their mistakes more effectively than in traditional class settings. The relatively low standard deviation further indicates consistency across respondents.

4.3.2. Boost confidence in speaking

The item *"I felt more confident speaking English after using the app"* yielded a high score (M = 4.38, SD = 0.62), reflecting a strong psychological impact of ASR use. This finding is particularly significant in contexts where learners often experience anxiety or self-consciousness about their pronunciation. Students likely benefited from the ability to practice in private, repeat difficult words, and receive feedback without fear of judgment.

4.3.3. Clarity and usefulness of ASR feedback

Students agreed that *"The app's feedback was timely and easy to understand"* (M = 4.46, SD = 0.55), reinforcing the pedagogical value of real-time corrective input. The automated nature of the feedback likely offered a more consistent and objective assessment than occasional teacher corrections in large classes.

4.3.4. Increased motivation and engagement

The item *"I was more motivated to practice pronunciation using the app"* scored M = 4.29 (SD = 0.63). This suggests that the interactive, game-like environment of ELSA Speak—featuring progress scores, achievement levels, and streak tracking—contributed positively to learner engagement. Students reported that these elements encouraged regular practice, which is essential for improving phonological accuracy.

4.3.5. Preference for ASR over traditional drills

While the item *"I prefer using ELSA Speak to traditional classroom drills"* received a slightly lower mean score (M = 4.12, SD = 0.73), the result still indicates a general preference for technology-enhanced instruction. The higher standard deviation suggests that some students may still value face-to-face interaction or have found the app less compatible with their learning style.

4.3.6. Continued use after course completion

Importantly, the item *"I will continue to use the app even after this course ends"* received a high score (M = 4.47, SD = 0.59), demonstrating strong learner autonomy and long-term interest in self-directed pronunciation improvement. This sustained engagement is a promising indicator of the app's usefulness beyond the classroom setting.

V. Discussion

The findings of this study offer compelling evidence that the integration of automatic speech recognition (ASR) technology, specifically the ELSA Speak application, can significantly enhance the pronunciation competence of EFL students in a Vietnamese university context. The improvements observed in segmental accuracy, suprasegmental features, and intelligibility were not only statistically significant but also pedagogically meaningful. This section discusses the implications of these results in light of existing literature and theoretical frameworks and considers both the strengths and limitations of ASR-assisted pronunciation instruction.

5.1. ASR technology and pronunciation gains

The experimental group demonstrated substantial improvements in all measured pronunciation components compared to the control group. These results are consistent with prior studies such as McCrocklin (2016), Elimat and AbuSeileek (2014), and Nguyen et al. (2025), which also reported significant gains in learners' phonetic accuracy through the use of ASR tools. Notably, the largest gains in this study were recorded in overall intelligibility, a finding that aligns with Derwing and Munro's (2015) argument that intelligibility is a more practical and achievable goal than native-like pronunciation in EFL contexts.

The fine-grained, real-time feedback provided by ELSA Speak likely played a crucial role in raising learners' phonological awareness. Unlike traditional classroom drills, which often rely on teacher modeling and delayed feedback, ASR systems allow learners to instantly detect and correct their mistakes. This mechanism supports a shift from teacher-centered to learner-centered instruction, enabling students to take ownership of their pronunciation practice.

5.2. Learner perceptions and motivation

Survey data indicated a high level of learner satisfaction with ASR-based instruction. Students appreciated the immediacy and clarity of feedback, reported increased confidence in speaking, and expressed strong intentions to continue using the application beyond the study period. These findings reinforce Vygotsky's sociocultural theory, which posits that learning is most effective when learners engage actively and receive timely support within their zone of proximal development. In this case, ASR technology acted as a mediating tool, bridging the gap between learners' current competence and their desired pronunciation goals.

The motivational aspect of ELSA Speak also deserves attention. As highlighted in gamification literature (e.g., Jayalath & Esichaikul, 2022), features such as badges, progress tracking, and levels can increase learner persistence and engagement. This is particularly valuable in pronunciation training, which is often repetitive and anxiety-inducing. By creating a low-pressure, self-paced environment, the app helped students sustain regular practice and develop pronunciation habits.

5.3. Pedagogical implications

The success of ASR integration in this study highlights several implications for language educators and curriculum designers:

- Blended Approaches: ASR should be viewed not as a replacement for classroom instruction, but as a complementary tool that enhances learners' access to feedback and practice time. Teachers can use ASR apps to supplement in-class activities and encourage self-study.

- Feedback Orientation: The ability of ASR tools to provide detailed, individualized feedback aligns well with formative assessment principles. Teachers should guide students in interpreting and acting on ASR feedback for deeper learning.

- Teacher Training: For successful implementation, instructors must be equipped with the digital literacy and pedagogical strategies necessary to integrate ASR meaningfully into their courses.

- Accessibility and Equity: Given the increasing affordability of mobile devices and internet access, ASR-based learning may serve as an equitable solution for pronunciation training in public universities with limited resources.

5.4. Limitations and future directions

While the findings are promising, several limitations should be acknowledged. First, the intervention period was relatively short (six weeks), and long-term retention of pronunciation gains was not assessed. Second, the study involved a specific group of non-English majors at a single university, which may limit generalizability. Third, technical issues such as device compatibility, internet connectivity, and accent bias in ASR algorithms could affect the consistency of feedback.

Future research should explore the longitudinal effects of ASR use, its impact on spontaneous speech production, and how ASR tools perform across different learner proficiency levels and sociolinguistic backgrounds. Comparative studies between different ASR applications could also provide insight into which features most effectively support pronunciation learning.

VI. Conclusion

This study set out to examine the effectiveness of automatic speech recognition (ASR) technology, specifically the ELSA Speak mobile application, in improving the English pronunciation of non-English major students at the University of Labour and Social Affairs. Through a quasi-experimental design involving pre- and post-tests, supported by a learner perception survey, the study revealed that students who used ASR tools demonstrated significantly greater improvements in segmental accuracy, suprasegmental features, and overall intelligibility compared to those who followed traditional instruction alone.

The findings affirm the pedagogical value of integrating ASR into pronunciation instruction, particularly in EFL contexts where class sizes are large and personalized oral feedback is limited. The ability of ASR applications to deliver immediate, individualized, and consistent feedback empowers learners to take control of their own progress. Furthermore, positive student perceptions indicate strong motivation, increased confidence, and a willingness to continue using the technology beyond formal instruction, factors which are crucial for long-term skill development.

From a broader perspective, the study contributes to the growing body of evidence supporting the use of educational technologies to promote learner autonomy and personalized learning. It also underscores the importance of aligning digital tools with communicative learning goals and training teachers to incorporate such tools effectively into curriculum design.

However, the study's scope was limited in duration and sample size and did not assess long-term pronunciation retention or spontaneous speech performance. Future research should address these limitations, explore the integration of ASR in blended and flipped learning environments, and investigate how different learner profiles respond to ASR-based interventions.

In conclusion, ASR technology offers a promising and scalable solution to persistent challenges in English pronunciation instruction. When implemented thoughtfully and supported by pedagogical scaffolding, tools like ELSA Speak can enhance not only learners' pronunciation competence but also their confidence, motivation, and overall communicative effectiveness in English.

References

[1]. 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

[2]. Chapelle, C., & Jamieson, J. (2008). *Tips for teaching with CALL: Practical*

approaches to computer-assisted language learning. Pearson Education.

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

[4]. Elimat, A. K., & AbuSeileek, A. F. (2014). Automatic speech recognition technology as an effective means for teaching pronunciation. *JALT CALL Journal, 10*(1), 21-47. https://doi.org/10.29140/jaltcall.v10n1.166

[5]. Jayalath, J., & Esichaikul, V. (2022). Gamification to enhance motivation and engagement in blended eLearning for technical and vocational education and training. *Technology, Knowledge and Learning, 27*, 91-118. https://doi.org/10.1007/s10758-020-09466-2

[6]. 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

[7]. Nguyen, T. S., Nguyen, T. D. T., Hoang, N. Q. N., & Do, T. K. H. (2025). How AI-powered voice recognition has supported pronunciation competence among EFL university learners. *CALL-EJ, 26*(3), 64-83. https://doi.org/10.54855/callej.252634

[8]. Nguyen, L. T., & Newton, J. (2020). Pronunciation teaching in tertiary EFL classes: Vietnamese teachers' beliefs and practices. *TESL-EJ, 24*(1). https://www.tesl-ej.org/wordpress/issues/volume24/ej93/ej93a1/

[9]. Pham, V. T. T., & Pham, A. T. (2025). English major students' satisfaction with ELSA Speak in English pronunciation courses. *PLOS ONE, 20*(1), e0317378. https://doi.org/10.1371/journal.pone.0317378

[10]. Sholekhah, M. F., & Fakhrurriana, R. (2023). The use of ELSA Speak as a Mobile-Assisted Language Learning (MALL) towards EFL students' pronunciation. *JELITA, 2*(2), 93-100. https://doi.org/10.37058/jelita.v2i2.7596

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

ỨNG DỤNG CÔNG NGHỆ NHẬN DẠNG GIỌNG NÓI TRONG DẠY KỸ NĂNG NÓI CHO SINH VIÊN TRƯỜNG ĐẠI HỌC LAO ĐỘNG - XÃ HỘI

Lại Minh Thư[2]

Tóm tắt: *Nghiên cứu này nhằm đánh giá hiệu quả của việc ứng dụng công nghệ nhận dạng giọng nói (Automatic Speech Recognition - ASR), cụ thể là ứng dụng ELSA Speak, trong việc cải thiện kỹ năng phát âm tiếng Anh cho sinh viên không chuyên ngữ tại Trường Đại học Lao động - Xã hội. Nghiên cứu được triển khai theo thiết kế bán thực nghiệm với 60 sinh viên năm thứ nhất và thứ hai, chia thành hai nhóm: nhóm thực nghiệm sử dụng ELSA Speak trong 6 tuần và nhóm đối chứng học theo phương pháp truyền thống. Kết quả cho thấy nhóm thực nghiệm có sự cải thiện vượt trội về độ chính xác ngữ âm, các yếu tố siêu đoạn và mức độ dễ hiểu trong phát âm, với sự chênh lệch có ý nghĩa thống kê. Khảo sát cảm nhận người học cũng cho thấy sinh viên đánh giá cao tính hữu ích, dễ sử dụng và khả năng phản hồi kịp thời của ứng dụng. Các phát hiện khẳng định tiềm năng tích hợp ASR như một công cụ hỗ trợ hiệu quả cho việc giảng dạy phát âm trong bối cảnh đại học công lập tại Việt Nam, đồng thời gợi mở các hướng ứng dụng công nghệ trong giảng dạy ngôn ngữ theo hướng cá nhân hóa và tự học.*

Từ khóa: *phát âm tiếng Anh, nhận dạng giọng nói, ELSA Speak, công nghệ giáo dục, tự học, giảng dạy tiếng Anh*

[2] Trường Đại học Lao động - Xã hội