

Utilizing Artificial Intelligence Tools to Support Pronunciation Practice Among Non-English Major EFL Students

Tran Vu Mai Yen¹, Nguyen Tran Uyen Nhi^{1*}

¹ University of Foreign Language Studies, The University of Danang, Vietnam

*Corresponding author's email: ntunhi@ufl.udn.vn

 <https://orcid.org/0000-0003-4096-7673>

 <https://doi.org/10.54855/ijte.26645>

® Copyright (c) 2026 Tran Vu Mai Yen, Nguyen Tran Uyen Nhi

Received: 06/10/2026

Revision: 08/10/2026

Accepted: 17/10/2026

Online: 22/10/2026

ABSTRACT

Keywords: AI tool, EFL students, pronunciation practice, Speechify, Speechnotes, non-English majors

Although AI-powered speech technologies have shown promise for L2 pronunciation learning, little is known about the efficacy of combining complementary AI-powered tools for out-of-class pronunciation practice among Vietnamese non-English majors, particularly in terms of aligning AI-based error detection and teacher assessment. This study investigates the integration of Speechify and Speechnotes to support pronunciation self-practice among 35 first-year non-English major EFL students at a public university in Vietnam. Data were gathered through a descriptive survey (N=35) and semi-structured interviews (n=10). Results showed that students responded positively to the simultaneous use of these two tools, with all interviewed participants expressing satisfaction and 85,7% valuing Speechnotes' real-time visual feedback for error detection. Speechnotes' error-detection capability also showed partial alignment with teacher judgments in identifying segmental errors, including omitted final sounds and vowel contrasts. Along with future advancements in AI tool features, this study offers valuable insights to improve the quality of pronunciation instruction for non-English major students, foster greater engagement and motivation in pronunciation practice, and ultimately support foreign language learning.

Introduction

In a rapidly developing educational context, integrating technology into language learning is becoming increasingly essential. For non-English major EFL students, improving English pronunciation remains a significant challenge. One primary cause is the limited time allocated for pronunciation practice in the formal curriculum and the relatively large class sizes, which make it difficult for teachers to monitor and correct individual errors (Do, 2021). Additionally, some students feel anxious, shy, or lack confidence when practicing pronunciation, especially when their mistakes are noticed in front of the class.

Recent advances in artificial intelligence (AI) have opened up new avenues for language learning, offering dynamic, highly interactive, and personalized experiences. According to Godwin-Jones (2023), AI tools in language learning give learners opportunities to practice independently, receive immediate feedback, and gain data-driven insights (Nguyen & Nguyen, 2026). AI-supported tools, such as speech recognition software, virtual assistants, and language learning applications, are becoming increasingly popular due to their ability to recognize and analyze speech, detect pronunciation errors, provide instant feedback, and suggest specific corrections. Learners can practice anytime and anywhere, without being restricted by classroom space or the duration of face-to-face classes.

Given this practical context, investigating how AI tools can support pronunciation self-practice among non-English majors is essential. Therefore, this study evaluates the efficacy of two specific AI tools: Speechify (for listening and oral training) and Speechnotes (for speech recognition). These applications help first-year students detect phonetic errors and refine their pronunciation through automated, real-time feedback. Regular engagement with these digital resources not only supports continuous learning through pronunciation practice and accuracy verification but also fosters active, independent learning. Therefore, the study aims to investigate the advantages and challenges non-English major students encounter when using Speechify and Speechnotes for pronunciation self-practice, and to examine the extent to which Speechnotes' ability to detect pronunciation errors aligns with, or differs from, teachers' feedback. The findings are expected to enhance students' confidence, promote active participation, and improve their English pronunciation accuracy in both academic and practical environments.

Research Questions

To fulfill the purpose of the study, this research seeks to answer the following research questions:

1. What are the perceived advantages and challenges encountered by non-English major students when using Speechify and Speechnotes for pronunciation self-practice?
2. To what extent does Speechnotes' capacity in detecting pronunciation errors align with or differ from teachers' evaluations?

Literature Review

Pronunciation in Foreign Language Learning

For many EFL learners, English pronunciation is one of the most challenging skills, requiring significant time and persistent effort to improve. According to Celce-Murcia et al. (2010), pronunciation is not solely a technical aspect but a core component that demonstrates learners' language proficiency and is a critical factor in language teaching. For English learners, mastering clear and accurate pronunciation is essential to enable effective communication in both academic settings and daily interactions. With good pronunciation skills, learners can approach and utilize the language more effectively, particularly in productive oral tasks and listening comprehension. Conversely, weak pronunciation can pose a major barrier to

communication, leading to difficulties in the language learning process (Gilakjani, 2012). MacIntyre (1998) emphasizes that confidence in communication and the level of language anxiety are strongly influenced by learners' perceptions of their own pronunciation abilities. Being concerned about pronunciation errors, they tend to keep silent or avoid communication in real-life situations. Employers may highly value candidates with good pronunciation skills because they can communicate effectively with colleagues, clients, and stakeholders, which can lead to better career opportunities, prospects, and greater chances for advancement. Furthermore, pronunciation is not just a language skill but also reflects cultural identity and influences the interlocutor's perception. Using accurate pronunciation helps learners demonstrate cultural sensitivity and respect for the culture of native speakers, while enhancing communication effectiveness in multicultural environments (Derwing & Munro, 2005). Therefore, focusing on accurate and regular pronunciation practice will significantly help learners, especially non-English major students, improve their English capacity and expand future learning and career opportunities.

AI in Foreign Language Education

In recent years, AI has become a powerful tool in the field of language education, particularly within the EFL context. It offers learners many benefits, such as personalizing the learning process, enhancing interaction, providing instant feedback, and supporting learners in developing comprehensive language skills (Jegade, 2024; Qiao & Zhao, 2023). Regarding personalization, AI's ability to track and analyze each student's learning progress allows it to generate customized lessons based on individual needs and learning paces. AI can identify learners' strengths and weaknesses, helping them focus on weak areas and skip already mastered material. Online learning platforms utilizing AI can provide more effective language lessons by applying machine learning algorithms to track learning progress and adjust learning content to suit the learners' proficiency (Kabudi et al., 2021). Consequently, learners can experience an automated and personalized learning process to make language acquisition more efficient (Pham & Renandya, 2026).

In terms of creating an interactive learning environment, Huang et al. (2022) state that AI chatbots acting as virtual interactive agents provide a highly interactive setting where students practice communication flexibly and receive instant feedback, thereby enhancing learning efficiency. Additionally, in virtual reality environments, learners can interact with an AI assistant to practice language in real-world scenarios, helping to improve communication skills and cultural understanding of the target language.

Additionally, according to Kurniati and Fithriani (2022), AI-powered technologies in education have led to innovative teaching and learning experiences for teachers and students, including evaluation, tutoring, content generation, and automated feedback. Instant feedback and improved language skills are another major benefit, as AI provides immediate evaluations for pronunciation, intonation, writing correction, listening support, and vocabulary expansion. In pronunciation learning, technology-mediated tools such as YouGlish have also been reported to support learners' pronunciation accuracy, exposure to diverse accents, and autonomous learning (Nguyen & Nguyen, 2026). Using AI speech recognition technology (e.g., Google Speech, ELSA), the system analyzes pronunciation, intonation, and speech delivery speed,

helping learners speak more naturally and clearly. Thanks to instant feedback and continuous availability, AI helps learners recognize mistakes quickly, correct errors effectively, and develop comprehensive language skills, thereby increasing autonomy, efficiency, and positivity in the learning process (Godwin-Jones, 2023).

AI Tools for Pronunciation Practice

The support of AI tools in pronunciation practice is increasingly demonstrating its strengths in the academic environment for non-English major students who face limited class time and high levels of communicative anxiety (Le et al., 2024). Among currently popular tools, Speechify and Speechnotes can meet learners' needs thanks to their smart pronunciation support features, helping learners improve their pronunciation independently and flexibly (Figure 1 displays the interfaces of Speechify and Speechnotes)

Speechify

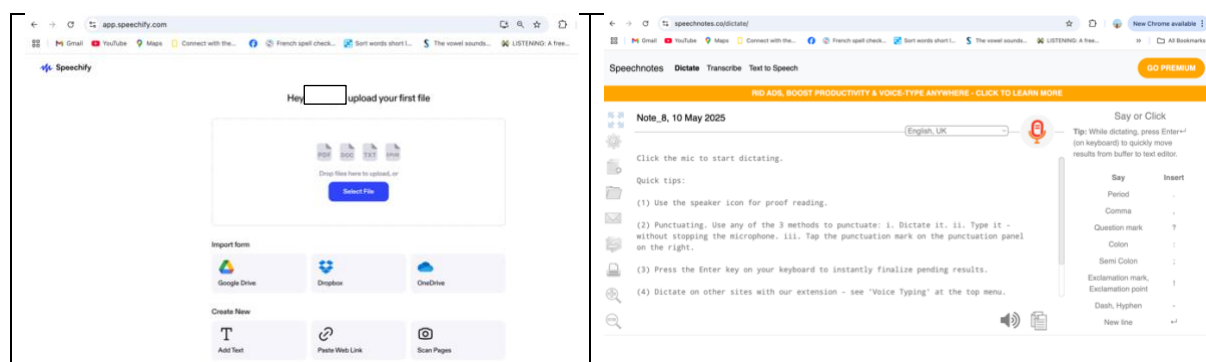
Speechify is a text-to-speech tool (TTS) that allows students to input or upload English documents, after which the system reads the text aloud using various native voices, such as American, British, Australian, etc. In computer-assisted pronunciation training (CAPT), Speechify serves as a high-quality tool for auditory input modeling. Speechify includes features such as: Voice options (users can choose from various native voices to match their learning goals); adjustable pronunciation speed (users can listen to detailed syllables or increase speed to practice listening reflexes); and pronunciation comparison and imitation (users can listen to a sentence, pause, and read it back to compare pronunciation, intonation, and pausing).

Speechnotes

Speechnotes is an Automatic Speech Recognition (ASR) tool for practicing pronunciation through recording and voice-to-text features, utilizing advanced speech recognition technology. With Speechnotes, learners experience instant speech-to-text conversion: if the learner pronounces correctly, the text appears accurately; if there is a pronunciation error, the displayed word will be misspelled or may even change the intended meaning of the word. Additionally, learners can practice speaking text blocks or sample sentences by reading an English passage and using Speechnotes to verify the level of accuracy.

Figure 1

User Interfaces of Speechify and Speechnotes



ASR Dictation versus Dedicated Automatic Pronunciation Assessment (APA)

Empirical research by Cámara-Arenas et al. (2023) identified a critical theoretical limitation: standard commercial ASR tools—such as the Google ASR technology integrated into Speechnotes—cannot simultaneously meet the intrinsic and derived criteria required to function as a specialized automated pronunciation assessment (APA) system for teaching speech segments. Unlike specialized APA systems designed to directly detect and diagnose phonetic errors, Speechnotes operates as a speech-to-text system, providing only indirect feedback in the form of text displaying transcription errors. It does not directly diagnose phonemes, stress, or intonation errors, nor does it provide explicit instructions on how to correct them. Instead, learners must notice mismatches between their intended utterance and the transcribed text on the screen.

The pedagogical distinctions between the various types of feedback provided by this technology have been clearly demonstrated in the large-scale meta-analysis conducted by Ngo et al. (2024). Ngo et al. (2024) found that while the general benefit of using ASR in speech instruction is moderate (Hedges's $d = 0.69$, representing a moderate-to-large effect size), there is considerable variation depending on learning strategies and feedback characteristics: (1) ASR systems providing direct error correction (explicit corrective feedback) achieve a very high level of effectiveness ($g = 0.86$, a large effect size), whereas systems providing subsequent feedback (indirect feedback)—such as dictation-based ASR programs (e.g., Speechnotes)—yield only moderate results ($g = 0.50$, a moderate effect size); (2) ASR tools significantly influence segmental feature development ($g = 0.82$) but have little effect on suprasegmental feature development ($g = 0.37$, a small effect size); and (3) ASR practice in a collaborative peer setting produces better results ($g = 0.89$) than independent practice ($g = 0.44$). These empirical findings provide an effective theoretical foundation for explaining learning mechanisms and shaping the accompanying supplementary activities when implementing Speechnotes for non-majors in Vietnam.

The combination of Speechify's input model and Speechnotes' indirect error-detection mechanism creates an effective, mutually reinforcing learning process. To ensure accurate machine recognition, students must actively improve their segmental pronunciation rather than passively receive information. However, the use of these tools requires clear pedagogical guidance from teachers to minimize undesirable coping strategies—such as students deliberately speaking extremely slowly to enhance recognition—which can unintentionally affect the fluency and natural rhythm of spoken language.

Research Gap and Study Context

Although prior studies have demonstrated the potential of ASR technology for the pronunciation skills of second language (L2) learners, significant research gaps remain. Most studies focus on the isolated application of a single technology (either TTS or ASR alone) within controlled classroom settings, rather than evaluating the simultaneous integration of two complementary self-study tools (Speechify and Speechnotes) in students' real-life contexts outside the classroom. In addition, there is a lack of empirical research evaluating the compatibility between machine-based pronunciation error detection (ASR dictation) and actual

teacher assessments regarding non-English major students in Vietnam- a group deeply influenced by negative L1 transfer when moving from a monosyllabic, tonal language to a polysyllabic language marked by stress and rhythm.

With advanced features, both tools benefit non-English major students by helping them enhance their pronunciation skills alongside formal class hours. These tools also help students form self-study habits, thereby increasing their confidence in communication. Consequently, this research aims to evaluate the effectiveness of Speechify and Speechnotes in improving pronunciation accuracy and error recognition, particularly regarding linking sounds and word stress. It also explores students' perceptions of the benefits and linguistic challenges, such as the influence of the Vietnamese language, they encounter when integrating these AI tools into their self-study routines.

Methods

Pedagogical Setting & Participants

The research was conducted at the University of Danang from February to April 2025, involving 35 first-year non-English major students (Aged 18–19; 63% female, 37% male) enrolled in a Pre-intermediate General English course. The participants were predominantly from Mechanical Engineering and Transportation and Road Engineering. 73.3% reported no prior experience with AI-assisted pronunciation tools. As part of the pedagogical framework, participants were required to integrate Speechify and Speechnotes into their self-study routines and submit weekly progress reports to the instructor. These reports documented the frequency of practice, total hours completed, accuracy rates, and specific pronunciation difficulties encountered. In addition, students were required to record their practice sessions weekly, and the teacher used the data to monitor practice frequency and provide appropriate feedback on both the practice content and recording techniques, ensuring the most accurate recognition by Speechnotes. The first author served as the course instructor, allowing for direct pedagogical observation; however, to mitigate power-relation bias, students were informed that participation in the research was voluntary and would not influence their final grades.

Design of the Study

This study adopts a qualitative-dominant mixed-methods classroom case study design, supplemented by a small-scale descriptive survey for triangulation. First, an initial descriptive survey with 35 students was conducted following their weekly pronunciation practice and prior exposure to AI-assisted pronunciation tools, using frequencies and percentages. Next, a purposive sampling technique was used to select 10 students for in-depth interviews. The selection criteria targeted highly engaged learners, defined as those who documented 10 hours or more of practice throughout the course. This sequential design allowed the qualitative insights from interviews to expand upon the initial survey tendencies, while a paired analysis of recordings specifically addressed the comparative requirements of RQ 2.

Research procedure

The 10-week intervention required participants to integrate Speechify and Speechnotes into their weekly self-study routines. Speechify was used for auditory modeling, with students instructed to select the “Natural” American or British English voice at a 1.0× speed setting for listening and imitation practice. Speechnotes was used as a speech-to-text dictation tool, with the language settings configured to American or British English. Participants were required to practice for a minimum of 60 minutes per week, focusing on specific phonological features, including segmentals (vowels and final consonants) and suprasegmentals (word stress and linking sounds). To reflect authentic self-study conditions, students recorded their practice sessions using personal mobile devices or laptops in settings outside classrooms.

Data Collection & Analysis

The study employed three primary data sources. First, a web-based Google Forms questionnaire comprised 15 items, including 5-point Likert-scale items and closed-ended questions designed to investigate students' current weekly pronunciation practice routines and their previous access to AI technology. The questionnaire items were adapted from previous research on mobile-assisted pronunciation learning and ASR-based feedback (Dai & Wu, 2023). To ensure face and content validity, the questionnaire was reviewed by an EFL lecturer with relevant expertise. One of the authors, an experienced EFL lecturer in translation, translated the questionnaire into Vietnamese to ensure linguistic accuracy and contextual appropriateness. All percentages reported in the study were calculated based on the full sample of 35 participants ($N = 35$). Second, purposive sampling was used to select 10 highly engaged students who had documented more than 10 hours of pronunciation practice throughout the course for semi-structured interviews (see Appendix 1). The reason for this choice is that participants engaging in continuous, in-depth interaction would encounter enough issues to identify Speechnotes' s algorithmic shortcomings, such as automatic vocabulary correction or the failure to identify suprasegmental features, while simultaneously fully utilizing the pedagogical potential of both Speechify and Speechnotes. Although this sampling strategy focused on more engaged users and may have introduced engagement bias, it provided rich insights into the tools' potential; the study acknowledges this limitation. The study analyzed qualitative data from the semi-structured interviews using Braun and Clarke's (2006) inductive thematic analysis framework, which involved verbatim transcription, initial line-by-line coding, grouping codes into larger thematic clusters, sequential reviewing, and defining final analytical categories.

Third, paired recording analysis was conducted to address RQ2 and examine the alignment between AI-generated and human judgments. The teacher evaluated a sample of student recordings using a pronunciation rubric based on the framework of Celce-Murcia et al. (2010), and the teacher's manual identification of errors was compared side-by-side with Speechnotes' transcription outputs to identify areas of agreement and divergence across segmental and suprasegmental features. The dataset consisted of 35 recordings of short talks from 35 participating students produced at the end of the course. Simultaneously, Speechnotes (an ASR engine) utilized the Google Speech-to-text algorithm to transcribe the audio recordings into text. This methodological improvement guarantees that statistical indicators, including the Positive agreement (P_{pos}) metric, reliably and fairly represent the device's true capacity to

detect pronunciation errors. This index visually reflects the percentage of instances where both Speechnotes and teachers reach absolute agreement when both identify the pronunciation as incorrect. Thus, the study employs the Positive agreement (P_{pos}) metric proposed by Cicchetti & Feinstein (1990) to avoid inflating the overall agreement rate because accurately pronounced instances constitute the majority of the data sample. Based on a 2 x 2 binary contingency matrix, the following mathematical formula for the P_{pos} index is established:

$$P_{pos} = \frac{2TP}{(2TP) + FP + FN} \times 100\%$$

Notes: TP- True positive (the number of instances where both the teacher and Speechnotes agreed in identifying the student's pronunciation errors)

FP- False Positive (the number of instances where the teacher assessed a student's pronunciation as correct, but Speechnotes recorded it as incorrect)

FN- False Negative (the number of instances where the teacher identified a pronunciation error, but Speechnotes recorded it as correct)

In order to most appropriately represent the AI's actual error-detection capabilities compared to humans, this formula only considers situations in which at least one party detects an error (within the scope of TP, FP, and FN).

Ethical Considerations

The study received ethics approval from the ESP Department, UFLS-UD. Informed consent was obtained from all participants before the intervention. To protect data privacy, we anonymized all audio recordings used for analysis and did not process any personally identifiable information through commercial AI platforms.

Findings and Discussion

The current situation regarding non-English majors' speaking practice and access to AI tools.

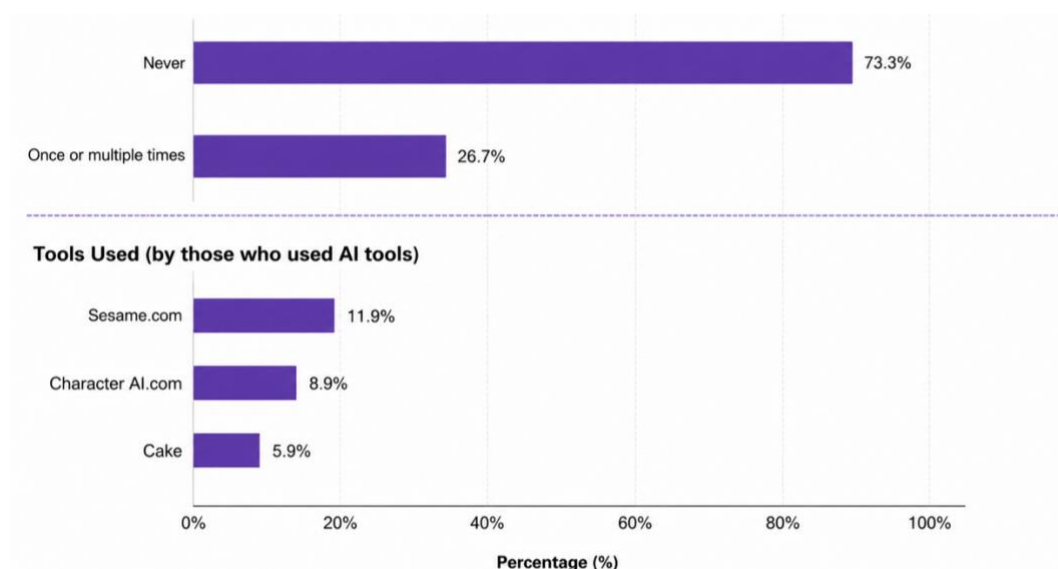
The data from 30 participants (N=35) revealed a clear and uneven difference in the amount of time spent practicing pronunciation. Notably, 60% (n=21/35) of the students surveyed spent less than 30 minutes per day on this activity. The group allocating between 30 minutes and one hour accounted for 31.4% (n=11/35), while only two participants invested more than one hour daily, and one student reported no regular practice. From the perspective of Second Language Acquisition (SLA) theory, the survey results indicate that the majority of learners are experiencing a severe shortage of opportunities to produce language output. Swain's output hypothesis (1985) emphasized that passive input is insufficient; learners must produce speech output to activate the automatic processes of grammar and phonology. The fact that 60% of students practice for less than 30 minutes a day highlights a significant obstacle in the English language learning environment (EFL context) in Vietnam, where students have little chance to practice real-world communication outside of the classroom.

The survey data also reflected the level of access to and use of artificial intelligence (AI) technology to improve English pronunciation skills among the research group. As can be seen

from Figure 2, 73.3% (n=26/35) of participants had no prior experience with AI-assisted pronunciation tools. This high figure reveals a significant gap in self-learning, as most students still rely on traditional methods or are unaware of the potential of modern technological applications. This aligns with Bala (2022), who noted that university students often lack self-direction and self-regulated learning skills outside of class, despite their desire for learning autonomy. On the other hand, just a small portion of students, 26.7% (n=9/35), had access to additional resources. Looking further into the categories of tools used, 5.9% of the participants stated that they were familiar with popular platforms like Cake and Sesame (11.9%) or intelligent conversational applications like Character.AI (8.9%). This finding shows that, although many AI-integrated programs support sophisticated speech recognition, most non-English majors' daily practice routines have not yet been significantly affected by this technological trend. This lack of practical experience underscores the need for targeted instruction and guidance so students can use AI tools to improve their pronunciation.

Figure 2

The reality of students' access to AI Tools



Overall, the severe lack of practice time and access to AI technology opened up opportunities for potential applications of ASR tools like Speechify and Speechnotes. By providing a low-pressure practice environment at home, AI tools act as a behavioral accelerator, encouraging students to significantly extend their interaction time and speaking frequency.

RQ1: What are the perceived advantages and challenges encountered by non-English major students when using Speechify and Speechnotes for pronunciation self-practice?

a. Advantages

The survey results revealed a generally positive attitude among students toward utilizing Speechify and Speechnotes for pronunciation self-practice. Despite most participants having no prior experience with AI-assisted language learning tools, both applications received high ratings for user-friendliness and accessibility. Regarding design, all interviewed students (10/10) were completely satisfied, highlighting their simple, clean, and intuitive interfaces.

These platforms offer easy-to-use interfaces suitable for various users, which can be operated easily and freely on personal computers or downloaded as mobile apps.

In terms of functional utility, Speechify was highly valued by 74.3% (n= 26/35) of the respondents. Students particularly appreciated its versatile feature set, which includes multi-voice selection (e.g., male, female, and native accents), adjustable playback speed with natural intonation, and extensive compatibility with multiple document formats (e.g., PDF, Word, and Google Docs). Similarly, Speechnotes was well-received, with 85.7% (n=30/35) of students agreeing that the tool possesses essential features for effective pronunciation training. Key advantages included the ability to play back and compare personal audio with the generated text, cross-situational practice (anywhere and anytime), immediate error detection, and a hassle-free experience with no registration required. Some specific responses are as follows:

"Every time I practiced pronouncing a word, a sentence, or a whole paragraph, I could see the text displayed on Speechnotes. If the displayed text does not match my recording, I would know that my pronunciation is not standard yet" (Student 02).

"When Speechnotes failed to recognize the word I said, I would pay closer attention to practising word stress and linking sounds" (Student 07).

Additionally, a deeper thematic analysis of student feedback indicated that Speechnotes served as a diagnostic tool, helping learners identify specific pronunciation errors. For instance, students noted difficulties in distinguishing short and long vowels and placing lexical stress:

"Speechnotes misrecognized words because I did not distinguish short and long vowels when pronouncing. For example: eat-it, a bit-beat, sport-spot" (Student 08).

"Speechnotes helped me detect that I misread word stress, such as 're'cord' instead of 'record'" (Student 07).

In addition to identifying their own pronunciation errors, the interview participants also evaluated their perceived improvement in their overall spoken delivery after extended practice with Speechify and Speechnotes. The responses were overwhelmingly positive, particularly regarding pronunciation accuracy and fluency. For the eight students who had no prior exposure to AI-assisted pronunciation tools, this intervention served as a critical foundation, significantly enhancing their ability to deliver relatively fluent presentations on familiar topics and better preparing them for institutional oral assessments. Conversely, the remaining two students, who had previously utilized applications such as Elsa Speak or the transcription feature of Google Translate, reported that integrating Speechify and Speechnotes into their existing routines offered a more diversified technological experience. This multi-tool approach allowed them to progressively tailor their pronunciation practice in alignment with their individual language learning pathways. The specific progress and trajectory of these two experienced individuals are illustrated in Figure 3.

Some noteworthy feedback regarding these improvements includes:

" I have noticed a significant improvement in my pronunciation. I can now pronounce more accurately and speak more fluently compared to before using these two tools" (Student 06).

" By repeatedly imitating the native speaker's voice on Speechify, I have improved my word stress and final sounds, while also focusing more on intonation when speaking." (Student 10).

"Previously, I was unaware that I frequently mispronounced word-final consonants or long-short vowels. Since practising with Speechnotes, I can easily pinpoint my mistakes and gradually rectify these errors." (Student 05).

Figure 3

Pronunciation improvement of two students (5 and 10) after practice sessions

Student 05's 6th sessions

Student 05's 8th sessions

Student 10's 27th sessions

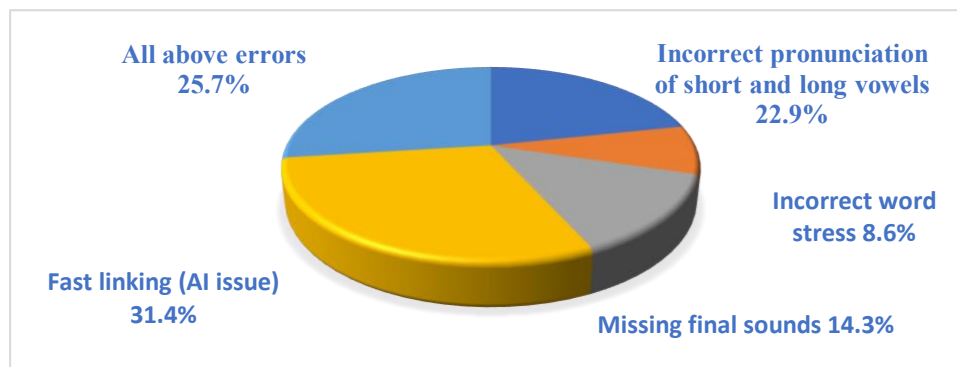
Student 10's 29th sessions

b. Challenges

For EFL learners, accurate pronunciation is often a primary barrier that leads to hesitation in language production. Such phonological difficulties frequently induce communicative anxiety and diminish confidence during interaction. Figure 4 below displays the percentage of pronunciation errors produced by students while practicing with Speechnotes.

Figure 4

Percentage of pronunciation errors when students use Speechnotes



The most common error, exhibited by 31.4% ($n=11/35$) of the students, involved fast sound-linking, which impeded the AI's ability to accurately recognize the users' intended utterances. This frequent issue likely stems from the L1 transfer of speaking rapidly and continuously, a common trait in Vietnamese speech patterns, or from delivery speeds that compromised the accuracy of the AI speech recognition technology. Additionally, 22.9% ($n=8/35$) of the participants reported difficulties with the mispronunciation of short and long vowels, underscoring a broader challenge in distinguishing and producing distinct English phonemes. The remaining infractions included the omission of final sounds (14.3%, $n=5/35$) and incorrect lexical stress placement (8.6%, $n=3$), both of which, despite lower frequencies, still significantly hindered the AI's processing capability. Notably, among the 35 surveyed students, nine (25.7%) demonstrated a combination of all the aforementioned pronunciation errors. As one student shared:

"There are some words that are difficult for me to pronounce correctly, such as the final sound in the word 'joints'; the app recognized the result as 'joyce', or when I linked sounds too quickly, the machine produced a different word, like 'his knee' into 'hickney'" (Student 04).

The cross-linguistic influence of the Vietnamese language also profoundly shapes students' phonological transfer when acquiring a foreign language. Due to the absence of certain English coda consonants in the Vietnamese phonetic inventory, learners frequently substitute complex English final sounds with familiar native equivalents. As noted by Nguyen (2018), "English is considered a prototypical stress-timed language, while Vietnamese is claimed to cluster with syllable-timed languages" (p.2). Consequently, Vietnamese speakers tend to apply an isosyllabic approach to English utterances, pronouncing each syllable with relatively equal intensity and pitch. This often leads to flat prosody, unaccented speech, or misplaced lexical stress, contrasting sharply with the variable stress placement inherent in English phonology.

These findings clearly indicate that student pronunciation challenges are largely rooted in interlinguistic differences between Vietnamese and English, combined with a lack of systematic practice and pedagogical feedback. In this regard, AI-driven applications serve as a helpful support tool by spotting pronunciation mistakes and delivering immediate corrective feedback. Quantitative and qualitative participant insights confirm that sustained engagement with these

tools mitigated persistent errors, as real-time feedback empowered students to consciously monitor, timely adjust, and systematically rectify their phonetic inaccuracies. Some specific interview feedback is as follows:

"A few words were mispronounced into other words not matching my intention, like more => mom, faster => facter. This might be because I haven't pronounced words in a standard way, making it hard for the machine to recognize. In subsequent times, I spoke very, very slowly, and the software recognized correctly" (Student 02).

"The software mainly recognized errors in ending sounds like 's' and 'es'. Basically, it's because the first time practicing pronunciation wasn't standard, making it hard for the machine to recognize. Through subsequent practice sessions, I found this error was significantly limited because I read slower and paused at short sentences. Even when I spoke multiple sentences, the app still fully recognized them with very few errors" (Student 08).

"Sometimes I tried to speak fluently but forgot to pay attention to my pronunciation, so Speechnotes would misrecognize words or record wrong content. After subsequent practice sessions, I gradually improved my pronunciation errors" (Student 06).

To mitigate the aforementioned challenges and optimize the efficacy of these applications, the participants were encouraged to offer constructive recommendations for future software enhancements. Specific pedagogical and technical suggestions shared by the students are as follows:

"Speechnotes should incorporate a real-time error detection feature that provides immediate phonetic transcriptions as a word is read. This would allow users to instantly recognize their mistakes and focus more effectively on corrective practice." (Student 04).

"Speechnotes should increase the word limit to 500 or 1000 words in the free version so that users can practice longer passages, which would better facilitate speaking practice for oral presentations." (Student 05).

"Speechify should introduce a 0.5 times playback speed option or a customizable mode that allows users to adjust the speed of individual segments. This would enable learners to clearly discern pronunciation details, lexical stress, and sound-linking patterns within a sentence." (Student 07).

"While both tools possess distinct advantages, integrating their core features into a unified platform would maximize learning outcomes. Ideally, a single application should allow users to upload text, listen to a native speaker's voice, and read along simultaneously, while automatically highlighting and flagging articulation errors for immediate tracking and correction." (Student 01).

Beyond these software-specific enhancements, several participants proposed supplementary strategies to augment their pronunciation training. Specifically, they suggested integrating Speechify and Speechnotes with external digital resources, such as the Cake application, Google Translate's transcription feature, or ChatGPT, to create a more comprehensive, multi-platform learning ecosystem.

Overall, the research findings demonstrate that students maintained a highly positive attitude toward using Speechify and Speechnotes for pronunciation self-practice, praising their user-friendly and accessible interfaces. Regarding specific advantages, Speechify was valued for its diverse voice options, adjustable speeds, and wide document compatibility, while Speechnotes stood out by allowing flexible, anytime practice with instant visual feedback that helped learners self-detect errors in word stress and vowel length. However, students still faced certain challenges. Pronunciation obstacles and systemic errors often triggered anxiety and diminished their confidence, while specific technical features, such as inserting punctuation via voice in Speechnotes, proved slightly difficult to navigate. Nevertheless, despite these barriers, students asserted that regular practice resulted in greater learner autonomy and confidence, with 8 out of 10 interviewed participants reporting significant improvements in their overall pronunciation accuracy and natural intonation.

RQ2: To what extent does Speechnotes' capacity in detecting pronunciation errors align with or differ from teachers' evaluations?

Oral speech input is decoded and transcribed using computational processes in automatic speech recognition. This is done by capturing oral voice input, analyzing it using a probabilistic algorithm, and producing an output, typically in the form of written text (Levis & Suvorov, 2012). Speechnotes operates based on a speech-to-text conversion algorithm. When students pronounce words, the system analyzes the mechanical properties of the sound waves and compares them to a probabilistic database to immediately display the result in the form of words or written text. If students mispronounce a word, Speechnotes doesn't directly catch the error but reflects it through incorrect typing or missing words, requiring the learner to deduce the pronunciation gap from the character errors on the screen. On the other hand, teachers employ a sophisticated cognitive network (rater cognition) to assess spoken language. They not only analyze mechanical and audio characteristics but also actively integrate contextual and communicative information and the learner's psychological factors at the time of speaking (Isaacs & Trofimovich, 2016). This depth of knowledge enables teachers to identify the reasons for students' systemic problems rather than merely seeing their surface effects. As a result, they can offer learners solutions that actually address the core of the problem, including showing correct pronunciation.

Throughout the experiment, the teacher also listened to the students' recordings and compared them with the material that Speechnotes had recognized in order to further illustrate how effective Speechnotes is at spotting pronunciation mistakes in students' self-practice before class and obtaining feedback from the teacher. Students' pronunciation skills are assessed using the two axes of segmentation and suprasegmentals, which are based on the traditional phonetics teaching theory of Celce-Murcia et al. (2010). According to Selinker (1972), language learners possess an intermediate language system (Interlanguage), which is a temporary phonetic system located between their mother tongue (L1) and the target language (L2). Additionally, Speechnotes is mostly used by non-language-major students for independent study in non-acoustically standardized settings like homes, cafes, or dorm rooms. If the technology has poor noise tolerance, transcription errors will be attributed to environmental noise rather than the student's pronunciation skill. Therefore, this aspect is considered solely as a practical factor

affecting the tool's implementation feasibility, rather than as a standard for acoustic agreement with teachers- who possess a superior, natural auditory filter capable of recognizing speech despite background noise. AI tools (like Speechnotes) are trained on large datasets of extensive spoken corpora and L1 data (Koenecke, 2020). When encountering an L2 accent, the system scores based on the algorithmic distance and often refuses to recognize it or misspells it (Markl, 2022). Conversely, human assessment strongly favors semantic intelligibility- helping learners effectively communicate in real-life situations, rather than imposing and forcing a mechanical, stereotypical standard of British-American English pronunciation (Levis, 2005). This systemic difference between the machine's mechanical lexical probabilistic model and the human's flexible communicative perception makes the L1 Interlanguage aspect impossible to quantify as a criterion for calculating statistical agreement rates. Thus, according to the theoretical framework of Celce-Murcia et al. (2010), the evaluation of agreement between the teacher and Speechnotes was restricted to two fundamental phonetic feature dimensions: segmental features (such as vowel contrasts and final consonants) and suprasegmental features (such as linking sounds and word stress). Table 1 below presents the agreement index between Speechnotes and teachers for each phonological category (N = 35).

Table 1.

Alignment analysis between Speechnotes (ASR) and teacher evaluations (N = 35)

Phonological category	Teacher -coded errors	Speechnotes's error recognition	True positive	False positive	False negative	P_pos
Vowel contrasts	15	10	10	0	5	80%
Final consonants	25	18	18	0	7	83,7%
Linking sounds	10	0	0	0	10	0.00%
Word stress	21	4	4	0	17	32%
Overall	71	32	32	0	39	62,1%

Note. TP= True Positive; FP= False Positive; FN= False Negative

A significant finding from the study is that, even when the student's pronunciation was completely accurate regarding the target phonemes and the teacher marked it as correct, Speechnotes could not identify correct words on the screen due to microphone interference, background noise in the self-study environment, or the limitations of the commercial ASR model in recognizing non-native speech. Under the closely controlled experimental conditions of this study, such technical errors caused by system interference were strictly managed, resulting in a False positive (FP) value of zero.

The experimental results are fully consistent with and provide direct evidence for the meta-analysis by Ngo et al. (2024): ASR technology is highly effective in addressing segmental errors but has a minimal impact on suprasegmental features. The results indicate a high level of agreement between Speechnotes and the teacher regarding vowel errors, achieving an independent agreement rate (P_pos) of 80%. Out of the 15 vowel errors identified by the teacher, Speechnotes successfully detected 10 instances (True positives = 10), clearly evidenced by corresponding spelling errors (e.g., "eat" /i:t/ being transcribed as "it" /ɪt/, or "beat" /bi:t/ as "bit" /bɪt/). The system missed only 5 errors involving syllables with slight phonemic deviations

(False negatives = 5). These results demonstrate the effectiveness of Speechnotes' acoustic analysis algorithm in vowel recognition, offering a reliable tool for independent vowel practice.

Regarding final consonants, the consistency rate for independent agreement (P_pos) is 83.7%. As Vietnamese lacks complex final consonant clusters, non-majors often tend to drop final consonants. The teacher identified 25 instances of final consonant omission across 35 recordings; Speechnotes successfully detected 18 of these (True positive = 18), evidenced by characteristic errors such as the omission of the fricative /s/ in "place" (resulting in "play") or the loss of the stop /d/ in "friend" (resulting in "frien"). However, the tool missed 7 errors (False negative = 7) because the system automatically made up for the missing information by predicting the context based on surrounding words.

Notably, regarding the phenomenon of connected speech (specifically linking), the agreement rate was 0%, reflecting a level of poor agreement. Teachers observed 10 instances where students produced completely disjointed speech without linking (e.g., pronouncing "save energy" as two separate units—/'seɪv/ /'en.ə.dʒi/—instead of the linked /seɪ'ven.ə.dʒi/). However, Speechnotes recorded zero errors (FN = 10), meaning the system achieved 100% transcription accuracy regardless of whether the students employed connected speech. This occurs because Speechnotes operates using language models designed to optimize the fluency of the transcribed text. Consequently, this ASR tool is entirely unsuitable for detecting or assessing suprasegmental errors related to connected speech.

Finally, word stress (P_pos = 32%) achieved a low level of agreement. Teachers identified 21 utterances with incorrect word stress (such as misplacing stress in multisyllabic words or noun-verb pairs like "re'cord" and "'record"). Speechnotes accurately reflected 4 of these errors (TP = 4) by displaying the incorrect word. It simply converts audio into text based on lexical probabilities. Consequently, Speechnotes can only signal a stress error indirectly—specifically when the error is severe enough to cause the system to misidentify the word. In all other instances, the system automatically corrects the context to display the right word, thereby unintentionally ignoring the students' stress errors.

The research findings align with the study by Cámara-Arenas et al. (2023), which highlights the limitations of technology. Experiments with the suggested AI tools demonstrate that Speechnotes— a commercial speech-to-text tool— provides only indirect feedback based on transcription errors rather than functioning as a dedicated automated pronunciation assessment (APA) system. It cannot simultaneously meet the acoustic requirements for diagnosing suprasegmental errors in the manner of a specialized CAPT system. As a result, both teachers and students must be fully aware of the technology's limits. Speechnotes can serve as a useful adjunct for independent study or peer learning (Ngo et al., 2024) to help students identify and correct vowel and final-sound errors. However, regarding highly suprasegmental features such as linking sounds and word stress, direct pedagogical intervention and teacher feedback remain essential to ensure students develop practical communicative competence.

Conclusion

Integrating AI into EFL education offers promising potential to supplement traditional teaching models, shifting from conventional instruction to modernized, technology-driven frameworks that improve both learning efficiency and outcomes. ASR technologies like Speechnotes and TTS tools like Speechify serve as low-pressure, personalized practice environments for use outside the classroom. This combination fosters learner autonomy and increases practice frequency by providing native-speaker audio models and immediate visual feedback, enabling students to identify gaps in their own pronunciation. The small-scale survey and experiment indicate that the two AI tools' greatest potential lies in their ability to help identify segmental pronunciation errors. Speechnotes' real-time character feedback helps students enhance their self-monitoring of auditory perception, enabling them to detect and correct mechanical errors on their own, such as missing final sounds or the confusion of contrasting vowel pairs. However, consistent with empirical findings, commercial ASR tools have not yet demonstrated clear effectiveness in improving suprasegmental features, such as linking sounds or sentence stress. Thanks to algorithms that optimize transcribed fluency, the system often corrects errors and displays the correct words, even when students misplace stress or drop sounds in connected speech. As a result, the real value of these apps is found in their role as pedagogical support tools rather than in their ability to totally replace the teacher. Additionally, as commercial AI tools like Speechify and Speechnotes continue to evolve, they provide flexible options that learners can integrate into their personalized practice pathways. Overall, using Speechify and Speechnotes is an effective strategy for reducing language-related anxiety and boosting confidence among non-English major students. Given the limited in-class practice time from large class sizes, these digital tools help students build a sustainable habit of self-monitoring their pronunciation, gradually improving accuracy over long-term independent study.

To maximize the efficacy and broader adoption of AI-assisted pronunciation tools for non-English majors, several pedagogical and practical recommendations are proposed for both learners and instructors. For students, active engagement in self-monitoring and self-evaluation is critical to fostering autonomous learning motivation across both classroom and independent environments. Additionally, learners are encouraged to engage in collaborative knowledge-sharing by exchanging personal experiences with AI tools and providing peer feedback based on Speechnotes recordings. Furthermore, integrating Speechify for input modeling alongside Speechnotes for recording group presentations can significantly enhance collaborative speaking competencies. For instructors, it is essential to seamlessly integrate AI-supported speaking sessions as a formal component of the instructional curriculum. Consequently, instructors must have the digital literacy to guide students in using appropriate AI applications aligned with specific instructional objectives. Finally, establishing a continuous feedback loop from both students and colleagues during implementation will facilitate timely pedagogical adjustments. This iterative process not only enhances overall teaching efficiency and quality but also promotes the development of interactive forums dedicated to sharing best practices for classroom AI integration.

Limitations and Future Research

Although this study has provided significant practical insights into integrating the AI tools Speechify and Speechnotes to improve students' self-study pronunciation, the current methodological design still presents certain limitations that need to be acknowledged. First, a limitation of the study lies in its relatively small and localized sample size. The self-study experiment and quantitative survey were conducted with a limited group of 35 non-English major students at a public university in Vietnam. Furthermore, the qualitative process involved in-depth interviews focused on a small group of the 10 most active participants (those with over 10 hours of practice per course) to fully explore the effectiveness of the two AI tools; consequently, the findings cannot be generalized to the entire population of non-English major students in Vietnam. Secondly, the study focused solely on evaluating the feasibility of two AI tools in supporting students' pronunciation practice based on surveys on students' perceptions and teachers' analysis of their recordings but lacked the use of standardized pre- and post-tests or a control group.

To address these limitations, future research should be expanded to include multiple classes, courses, and educational institutions with larger sample sizes. This would enhance the generalizability of the data and enable a more thorough analysis of how Speechnotes and Speechify affect student groups with different English levels. Furthermore, future research should consider objective pre- and post-intervention tests. This approach, combined with the application of agreement metrics (such as Cohen's Kappa calculated by multiple independent raters), will provide quantitative evidence to confirm whether the AI-integrated self-learning model performs better than conventional pedagogical methods.

References

- Bala, E. (2022). The Perceptions and Practices of Learner Autonomy among Undergraduate English as a Foreign Language Students in Private Universities (A Case of Erbil/Iraq). *International Journal of Social Sciences & Educational Studies*, 9(1), 369-382. <https://doi.org/10.23918/ijsses.v9i1p369>.
- Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77-101. <https://doi.org/10.1191/1478088706qp063oa>
- Cámara-Arenas, E., Tejedor-García, C., Tomas-Vázquez, C. J., & Escudero-Mancebo, D. (2023). Automatic pronunciation assessment vs. automatic speech recognition: A study of conflicting conditions for L2 English. *Language Learning & Technology*, 27(1), 1-19. <https://doi.org/10.64152/10125/73512>
- Celce-Murcia, M., Brinton, D. M., & Goodwin, J. M. (2010). *Teaching pronunciation: A course book and reference guide* (2nd ed.). Cambridge University Press.
- Cicchetti, D. V., & Feinstein, A. R. (1990). High agreement but low kappa: II. Resolving the paradoxes. *Journal of Clinical Epidemiology*, 43(6), 551-558. [https://doi.org/10.1016/0895-4356\(90\)90159-M](https://doi.org/10.1016/0895-4356(90)90159-M)
- 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>
- Derwing, T. M., & Munro, M. J. (2005). Second language accent and pronunciation teaching: A research-based approach. *TESOL Quarterly*, 39(3), 379-397. <https://doi.org/10.2307/3588486>

- Do, L. M. (2021). The role of pronunciation teaching in EFL classrooms: Beliefs and practices of teachers and students. *International Journal of TESOL & Education*, 1(1), 47–53. <https://i-jte.org/index.php/journal/article/view/7>
- Gilakjani, A. P. (2012). *The significance of pronunciation in English language teaching*. *English Language Teaching*, 5(4), 96–107. <https://doi.org/10.5539/elt.v5n4p96>
- Godwin-Jones, R. (2023). *AI in language learning: Opportunities and threats* [Short paper]. UNAD Conference, Colombia. https://godwinjones.com/godwin-jones_AI_in_language_learning.pdf
- Huang, H., Hew, K. F., & Fryer, L. K. (2022). Chatbots for language learning—Are they really useful? A systematic review of chatbot-supported language learning. *Journal of Computer Assisted Learning*, 38(1), 237–257. <https://doi.org/10.1111/jcal.12610>
- Isaacs, T., & Trofimovich, P. (Eds.). (2016). *Second language pronunciation assessment: Interdisciplinary perspectives*. Multilingual Matters.
- Jegede, O. O. (2024). Artificial intelligence and English language learning: Exploring the roles of AI-driven tools in personalizing learning and providing instant feedback. *Universal Library of Languages and Literatures*, 1(2), 6–19. <https://doi.org/10.70315/uloap.ullli.2024.0102002>
- Kabudi, T., Pappas, I., & Olsen, D. H. (2021). AI-enabled adaptive learning systems: A systematic mapping of the literature. *Computers and Education: Artificial Intelligence*, 2, Article 100017. <https://doi.org/10.1016/j.caeai.2021.100017>
- A. Koenecke, A. Nam, E. Lake, J. Nudell, M. Quartey, Z. Mengesha, C. Toups, J.R. Rickford, D. Jurafsky, & S. Goel. (2020). Racial disparities in automated speech recognition, *Proc. Natl. Acad. Sci. U.S.A.* 117 (14) 7684–7689, <https://doi.org/10.1073/pnas.1915768117>.
- Kurniati, E. Y., & Fithriani, R. (2022). Post-graduate students' perceptions of QuillBot utilization in English academic writing class. *Journal of English Language Teaching and Linguistics*, 7(3), 437–448. <https://doi.org/10.21462/jeltl.v7i3.852>
- Levis, J. M. (2005). Changing contexts and shifting paradigms in pronunciation teaching. *TESOL Quarterly*, 39(3), 369–377. <https://doi.org/10.2307/3588485>
- Levis, J., & Suvorov, R. (2012). Automatic speech recognition. In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics* (pp. 1–8). Wiley-Blackwell. <https://doi.org/10.1002/9781405198431.wbeal0066>.
- Macintyre, P.D., Clément, R., Dornyel, Z. and Noels, K.A. (1998), Conceptualizing Willingness to Communicate in a L2: A Situational Model of L2 Confidence and Affiliation. *The Modern Language Journal*, 82: 545-562. <https://doi.org/10.1111/j.1540-4781.1998.tb05543.x>
- Markl, N. (2022). Language variation and algorithmic bias: Understanding algorithmic bias in British English automatic speech recognition. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 521–534). Association for Computing Machinery. <https://doi.org/10.1145/3531146.3533117>
- 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>.

- Nguyen, A.T. T. (2018). L2 English rhythm by Vietnamese speakers: A rhythm metric study. *The Linguistics Journal*, 12(1), 22–44.
- Nguyen, M. N., & Nguyen, T. M. H. (2026). YouGlish as a tool for enhancing English pronunciation: Students' perceptions at a university in Hanoi. *International Journal of TESOL & Education*, 6(2), 30–41. <https://doi.org/10.54855/ijte.26623>
- Pham, V. P. H., & Renandya, W. A. (2026). Navigating the AI turn: Framework for responsible AI integration in language education. *International Journal of TESOL & Education*, 6(2), 70–85. <https://doi.org/10.54855/ijte.26626>
- Qiao, H., & Zhao, A. (2023). Artificial intelligence-based language learning: Illuminating the impact on speaking skills and self-regulation in Chinese EFL context. *Frontiers in Psychology*, 14, <https://doi.org/10.3389/fpsyg.2023.1255594>.
- Selinker, L. (1972). Interlanguage. *International Review of Applied Linguistics in Language Teaching*, 10(1-4), 209-232. <https://doi.org/10.1515/iral.1972.10.1-4.209>
- Swain, M. (1985). Communicative competence: Some roles of comprehensible input and comprehensible output in its development. In S. Gass & C. Madden (Eds.), *Input in second language acquisition* (pp. 235–253). Newbury House.

Biodata

Tran Vu Mai Yen, M.A., is a lecturer of English at the University of Foreign Language Studies, the University of Da Nang, Da Nang City, Vietnam. Her primary research interests and areas of expertise include classroom pedagogy and integrating technology into English language teaching. She has published her research in journals such as the Journal of Science and Technology, Asiaccall Online Journal, and ESP Asian Journal.

Dr. Nguyen Tran Uyen Nhi is a lecturer of English at the ESP Faculty, University of Foreign Language Studies-University of Danang (UFL-UD), Vietnam. Her instructional design and research focus on Business English, Interpreting, CLIL, scaffolding strategies, and technology methodology. She actively contributes to the global TESOL and education communities through various scholarly activities, curriculum development projects, and research publications.

Appendice

Appendix 1: Interview questions

1. What difficulties did you encounter when using these two tools?
2. What pronunciation errors did Speechnotes detect?
3. Comment on the improvement in your speaking skills compared to before you used these two tools.
4. What suggestions do you have for improving pronunciation practice using these two tools?