IJIRSS
International Journal of Innovative Research & Scientific Studies

ISSN: 2617-6548

URL: www.ijirss.com

International Journal of Innovative Research and Scientific Studies
ONLINE ISSN: 2617-6548

AI-mediated pronunciation training: Vietnamese EFL learners' perceptions of ELSA speak

Vuong Thi Hai Yen[1*], Nguyen Thi Thu Huyen[2]

[1,2]*Faculty of Foreign Languages, Hanoi Metropolitan University, Vietnam.*

Corresponding author: Vuong Thi Hai Yen (*Email:* vthyen@hnmu.edu.vn)

Abstract

This study examines how first-year English majors at Hanoi Metropolitan University, Vietnam perceive ELSA Speak as a pronunciation learning tool, with a focus on learner autonomy, technological affordances, and institutional constraints within mobile-assisted language learning (MALL). Using a convergent parallel mixed-method design, the study collected quantitative data from 110 participants through a structured survey (Cronbach's alpha = 0.87) and qualitative data from semi-structured interviews with 24 purposively selected participants. Data were analysed using SPSS 26.0, thematic analysis following Braun and Clarke [1] and NVivo 12, within a theoretical framework integrating the Technology Acceptance Model and Self-Determination Theory. 77.3% percent of the participants found ELSA Speak effective or highly effective for improving their pronunciation. Learners reported increased confidence (81.8%) and motivation (68.2%), and they considered instant phoneme feedback the most valuable feature (M = 4.2, SD = 0.7). Three primary constraints were consistently cited: inadequate contextual practice (forty point nine percent), the expense associated with premium functionalities (36.4%), and connectivity issues (31.8%). Qualitative examination identified accent bias within speech recognition as a prevalent concern, specifically impacting the precision of feedback for English spoken with a Vietnamese accent. These observations indicate that ELSA Speak facilitates pronunciation practice and learner autonomy, despite accent bias, an over-dependence on automated feedback, and a restricted emphasis on suprasegmental features representing notable limitations. Therefore, educators should use ELSA Speak in blended learning environments that combine AI tools with traditional teaching methods. At the same time, institutions should address accessibility issues related to cost and infrastructure.

Keywords: AI-mediated training, ELSA Speak, Mobile-assisted language learning, Pronunciation training, Vietnamese EFL learners.

DOI: 10.53894/ijirss.v9i3.11309
Funding: This study received no specific financial support.
History: Received: 10 December 2025 / **Revised:** 4 February 2026 / **Accepted:** 9 February 2026 / **Published:** 3 March 2026

Competing Interests: The authors declare that they have no competing interests.
Authors' Contributions: All authors contributed equally to the conception and design of the study. All authors have read and agreed to the published version of the manuscript.
Transparency: The authors confirm that the manuscript is an honest, accurate, and transparent account of the study; that no vital features of the study have been omitted; and that any discrepancies from the study as planned have been explained. This study followed all ethical practices during writing.
Publisher: Innovative Research Publishing

1. Introduction

For learners of English in general and Vietnamese ones in particular, correct pronunciation play an important role in having effective and successful communication, particularly because of the significant differences between the phonological systems of Vietnamese and English English [2, 3]. One of the most challenges Vietnamese learners consistently face with are English phonemes absent in their native language, such as interdental fricatives /θ/ and /ð/, the postalveolar fricative /ʃ/, and complex consonant clusters [2]. Another problem is Vietnam's education testing system such as university entrance examination that focuses on tests and written skills over spoken skills [4] which makes these pronunciation problems even worse and worse. This is because formal classrooms fail to instruct pronunciation very well.

The advent of mobile-assisted language learning (MALL) technologies has opened up new ways to help individuals who have pronunciation problems by providing them personalized, easy-to-access, and apparently fun ways to learn [5-7]. ELSA (English Language Speech Assistant) Speak becomes a prominent application attracting learners of English. It uses AI and advanced speech recognition to give learners personalized, real-time feedback on how to pronounce words naturally and correctly [8]. But we need to think critically about this excitement: do AI-powered pronunciation apps really help learners, or do they mostly respond to market forces and technological determinism?

Recent studies on MALL applications show that students have positive attitudes towards them in enabling them get better in pronunciation. However, they also show three major challenges that this study wants to address. First, the findings demonstrate an uncritical faith in technology, in which often interprets learner enthusiasm as proof of effectiveness whereas overlooks the intrinsic limitations and associated risks of technology. Studies often highlight higher user satisfaction and motivation (e.g., [8, 9]) without carefully assessing whether these emotional advantages lead to sustained pronunciation enhancement or merely reflect short-term impacts and the temporary appeal from the game.

Second, previous study inadequately examines the socioeconomic and linguistic challenges of AI voice recognition technologies. Most voice recognition software is spoken by native speakers such as British and North American English [10]. Consequently, these systems illustrate ongoing accent bias, as proven by reduced accuracy and increased mistake rates in the processing of outer-circle and expanding-circle types, including Vietnamese-accented English [11]. This technological bias brings up important questions about whether AI applications are appropriate for cases in which the speaker is not a native speaker or not.

Third, previous research does not sufficiently concentrate on higher education settings and English major students. Vietnamese universities have a unique group of English majors having different educational needs, ways to improve their skills, and career goals than other EFL learners. Their purposes of pronouncing English accurately are not only for daily interactions but also for job opportunities such as teaching, translating, and interacting with other people from other countries, where comprehension is crucial.

This study aims to examine how first-year English majors at Hanoi Metropolitan University in Vietnam view ELSA Speak as a tool for improving their pronunciation skills. It will also evaluate the program's advantages and disadvantages within the context of mobile-assisted language learning (MALL).This study fills in the gaps left by earlier research by looking at how first-year English majors perceive ELSA Speak through a framework that have combinations with the Technology Acceptance Model (TAM; Davis [12]), Self-Determination Theory (SDT; Deci and Ryan [13]), and critical analyses of biases in algorithms implemented in AI systems. It also explores the challenges and difficulties faced by the learners when people learn how to pronounce words with the support of AI, instead of just asking them what they like about ELSA Speak. The results show that the challenge presented narratives about how technology can help with instructional issues. This led to the development of advanced, adaptive techniques that maximize the real benefits of AI applications while acknowledging and addressing their fundamental constraints.

1.1. Research Questions

To fulfill the purpose of the study, the survey was seeking to answer the following research questions

RQ1: How do first-year English majors at Hanoi Metropolitan University view the effectiveness of ELSA Speak in improving their pronunciation skills?

RQ2: What obstacles and constraints do learners face when using ELSA Speak for pronunciation practice?

RQ3: What are the best ways to use ELSA Speak in Vietnamese universities to help students learn pronunciation in English?

2. Literature Review

2.1. Mobile-Assisted Language Learning and Pronunciation

Pronunciation has significant influence on comprehension and communicative efficacy in second language acquisition [14, 15]. Over the last ten years, research in technology-enhanced language learning has seen a significant shift. It has moved away from models that focus on imitating native speakers, and instead, it now emphasizes communicative effectiveness and learner-centered design. Globally, Computer-Assisted Pronunciation Training (CAPT) systems, which use Automatic Speech Recognition (ASR), have become a major area of research. This is largely because they can provide immediate, personalized feedback on specific pronunciation features, such as how vowels and consonants are produced [16].

The emergence of generative AI tools has extended this further, with researchers examining how dialogue-based systems can simulate authentic conversation and reduce the foreign language anxiety that often limits speaking practice [17, 18].

MALL applications have already become useful resources to enhance pronunciation because they provide personalized practice opportunities, rapid feedback, and flexible environments for learning [19, 20]. Some theoretical frameworks such as constructivism and social learning theory that establish the pedagogical foundations of MALL focus on active learning, interactive experiences, and modeling in the construction of knowledge [21, 22]. Furthermore, artificial intelligence and speech recognition technologies are taken advantages of sending personalized feedback to learners and addressing specific pronunciation errors through adaptive learning pathways [17].

Many research studies have been conducted successfully to evaluate how MALL impacts on the instruction and acquisition of pronunciation. Lin and Lin [20] documented crucial improvements in learners' phonetic accuracy supported by mobile applications, which attributes these advancements to tailored practice opportunities independently. Brown and Ahmed [23] underscored the potential of adaptive learning strategies to adjust instruction and enhance individual skills and learning habits. They subsequently foster learners to get greater engagement and motivation. Interactive elements, including progress tracking, incentives, and challenges, have consistently shown considerable effectiveness in encouraging learners' involvement during repetitive pronunciation exercises [13, 24].

2.2. Critical Perspectives on AI-Powered Pronunciation Applications

The complex role of English Pronunciation Applications (EPA) in pronunciation teaching have been classified by recent meta-analyses and thorough reviews. Rogerson-Revell [16] important state-of-the-art review of Computer-Assisted Pronunciation Training (CAPT) proved that the immediate feedback of automatic speech recognition (ASR) produces reasonable effects to increase segments of pronunciation accuracy. Rogerson-Revell, on the other hand, pointed out that effectiveness varies greatly because it depends on the level of technology, the design of the lesson, and the complexity of the target feature. For example, suprasegmental aspects are harder to teach with computer technology than isolated phonemes.

There are a number of shortcomings with the current EPA research. Consequently, these drawbacks make it hard to use evidence-based practice. First, the methodology is still not very good because most studies use single-group pre-test/post-test designs without control groups. This makes it not easy to be sure that the advancements are due to EPA interventions and no other factors. Second, not enough attention is paid to the challenges caused by technology. Furthermore, ASR systems tend to favor certain accents all the time. For example, algorithms that are mostly trained on inner-circle English varieties are less accurate when they must deal with outer-circle and expanding-circle accents [10, 11]. This bias may show up as inconsistent feedback accuracy for Vietnamese learners, which could hurt their confidence and lead them to make mistakes when learning the way to pronounce words. Third, frameworks for pedagogical integration are still not fully developed. Important questions keep coming up including (1) What is the best balance between letting students use the EPA on their own and having a teacher guide them? (2) How can teachers make the most of EPA's strengths while making up for its weaknesses? Fourth, research incorrectly prioritizes segmental features over suprasegmental ones, despite substantial evidence demonstrating that stress, rhythm, and intonation are more critical to communication than phonemic accuracy [14].

2.3. The Application of Elsa Speak

Globalization and the expansion of digital tools have created new possibilities for language education. AI applications, in particular, offer students structured opportunities to practise speaking outside the classroom. Technology supports language acquisition by giving learners access to resources, enabling speaking practice, and facilitating interaction with native speakers [25]. Among the available tools, the English Language Speech Assistant (ELSA) application stands out for its interactive speaking exercises designed to actively engage learners in practising English.

ELSA Speak uses automatic speech recognition to provide instant feedback on pronunciation, intonation, and fluency. The application helps users practise consonant sounds, individual words, and full sentences through a mobile platform, allowing learners to study independently outside the classroom. Several studies support its effectiveness. Samad and Ismail [26] conducted a pre-experimental study with first-semester students at STKIP Muhammadiyah Enrekang and found that mean pronunciation scores rose from 1.96 to 5.79 after eight treatment sessions using ELSA Speak, with a t-value of 6.28 exceeding the critical value of 1.699 at the 0.05 significance level. More recent studies report similar positive outcomes in independent and mobile-based learning contexts [27-29]. Examining how this application works in practice may also inform broader decisions about digital tools in language education, with implications for both educators and curriculum developers [30].

Some research has already examined AI-based applications, including ELSA Speak, in the context of Indonesian students learning English [31, 32]. However, experimental studies that specifically test the effect of these tools on overall speaking proficiency remain limited, particularly in settings where English functions as the medium of instruction. Existing literature also identifies psychological barriers to speaking, such as anxiety and low confidence [33] yet empirical evidence on whether AI applications can address these barriers effectively remains scarce [34, 35]. A further gap concerns the specific role of immediate feedback: most studies treat technology use in language learning broadly, without isolating the effect of real-time AI feedback on speaking performance.

2.4. ELSA Speak in Vietnamese EFL Context

ELSA Speak is an AI-powered pronunciation application that uses automatic speech recognition (ASR) to deliver real-time feedback on individual phonemes and suprasegmental features such as stress and intonation [36]. Its use in Vietnam reflects a broader pattern of growing AI adoption in language education. Pham and Dang [37] report that approximately

86% of lecturers and over 91% of students at Vietnamese universities use AI tools in language teaching and learning, with ELSA Speak identified alongside ChatGPT, Grammarly, and other platforms as commonly used tools for language practice.

Research on ELSA Speak in Vietnamese contexts tends to report positive outcomes. Learners show measurable gains in recognising and correcting pronunciation errors, particularly with English phonemes absent from Vietnamese phonology [8]. Studies also report higher confidence and motivation, which researchers associate with the application's detailed feedback and progress-tracking features [9]. These findings align with the broader Vietnamese research trend showing that AI tools produce their strongest effects in assessment and feedback-related tasks, where automated scoring and performance analytics support self-monitoring [37, 38].

The phonological distance between Vietnamese and English creates specific learning demands that ELSA Speak may partially address. Vietnamese learners consistently struggle with English phonemes absent in their first language — such as interdental fricatives /θ/ and /ð/, final consonants, and consonant clusters — as well as with suprasegmental features including stress and rhythm [2, 3]. Research suggests that anchoring English pronunciation

Existing studies on ELSA Speak in Vietnam focus primarily on general learner populations and report satisfaction-based outcomes without examining higher education contexts or critically analysing the application's limitations from learners' perspectives. Infrastructure constraints — including unstable internet connectivity and unequal access to devices — also limit consistent use, particularly outside urban university settings [37]. This study addresses these gaps by examining the perceptions of first-year English majors at Hanoi Metropolitan University toward ELSA Speak, covering both its benefits and its limitations, to inform evidence-based technology integration in Vietnamese university EFL curricula.

3. Method

3.1. Research Design

This research employed a mixed-method design to gather and analyze data for an in-depth examination of participants' perceptions [39]. This methodology focuses on the perceptions of learners over quantifiable performance indicators, as outlined by the Technology Acceptance Model [12, 38] and Self-Determination Theory [7, 13]. The successful utilization and ongoing adoption of technology which are crucial for pedagogical effectiveness are wholly reliant upon respondents' subjective assessments of its utility, usability, and motivational effects.

The convergent design paid much attention to an comprehensive overview through the quantitative component, while intentionally utilizing qualitative data for more detail and background information. The analysis looked at these data sources in both positive and negative aspects get deeper insights.

3.2. Participants

110 in a total of 180 first-year English major students from Hanoi Metropolitan University, familiar with using ELSA Speak for pronunciation practice were chosen intentionally. In terms of the quantitative method, a convenient sampling was delivered to selected individuals with ELSA Speak experience, constituting 61.1% of the initial cohort. Demographically, 59.1% of the participants were female and 40.9% were male, with 90.9% being between the ages of 18 and 20. Self-reported levels of English proficiency were intermediate (77.3%), advanced (13.6%), and beginner (9.1%), which is representing typical incoming university English major profiles.

In terms of the qualitative method, intentional sampling identified 24 participants representing varied proficiency levels and engagement patterns with the application for ensuring information-rich cases for comprehensive analysis. Although the qualitative sample size is modest, it adheres to information power principles [40] and provides significant information power through precise objectives, purposive sampling, reliable theoretical frameworks, comprehensive interviews, and systematic thematic analysis.

3.3. Data Collection Instruments

3.3.1. Survey Questionnaire

A structured survey questionnaire with 25 questions assessed participants' perceptions across five dimensions: (1) perceived effectiveness in enhancing pronunciation, (2) learners' experience and interface design, (3) motivational impact, (4) learning autonomy and self-regulation, and (5) challenges and limitations. In the part of quantitative analysis, five-point Likert scales are used in some items, in which 1 means "Strongly Disagree" and 5 means "Strongly Agree respectively." Furthermore, there are demographic inquiries and open-ended questions to ask for experiences in details. The questionnaire showed satisfactory reliability (Cronbach's alpha = 0.87), which means that it was very consistent within itself. Three experienced TESOL practitioners reviewed the content and confirmed that the items in the survey questionnaire were clear, valid and relevant. The survey was given in English, which was the language of instruction for the participants, to make sure they understood it and gave honest answers.

3.3.2. Semi-Structured Interviews

The time for each semi-structured interviews lasted 25–30 minutes. Its purposes are to investigate deeply into participants' experiences. The format of the interview included 12 open-ended questions that were divided into 6 criteria such as (1) specific pronunciation improvements that ELSA Speak supported with, (2) the most helpful features for learning, (3) the challenges that came up, (5) how ELSA Speak compared to traditional teaching methods, and (6) suggestions for making it better. Follow-up questions encouraged elaboration and clarification, resulting in comprehensive and detailed evidence.

3.4. Data Collection Procedures

The data collection was conducted in the first semester of the 2025-2026 academic year. Before the data were gathered, full information including the purpose, processes, voluntary participation, and confidentiality was clearly given to all the participants. In addition, the respondents have the rights to give their informed consent or refuse to take part in the research. The survey was delivered to all respondents online in terms of Google Forms. This made it easy for them to access and manage the data. Whereas interviews were conducted in person in quiet, private settings, audio-recorded with the consent of participants, and subsequently transcribed in full for analysis.

3.5. Data Analysis

SPSS 26.0 was used to analyze the quantitative data. Descriptive statistics, including means, standard deviations, frequencies, and percentages, summarized participants' perceptions across survey question items. Associations between demographic variables and perception patterns where appropriate were measured with Chi-square tests.

Qualitative data were subjected to thematic analysis in accordance with Braun and Clarke [1] six-phase framework such as familiarization, initial coding, theme searching, theme reviewing, theme defining, and reporting. The analysis utilized both inductive and deductive techniques in which deductive coding applied theoretical frameworks (TAM, SDT), whereas inductive coding was responsive to new themes. NVivo 12 software is another useful data analysis instrument in the research because it makes code things easier and comes up with themes. To improve inter-coder reliability stronger, certain parts of the transcripts were coded carefully and double-checked.

4. Findings

4.1. Quantitative Findings

4.1.1. Overall Perceived Effectiveness

Most respondents who took the online survey said that ELSA Speak helped them improve their pronunciation. In particular, 77.3% of the respondents said that the app helped them improve their pronunciation skills or was extremely helpful. Only 4.5% thought it failed to be effective, and 18.2% reported to have neutral experiences. This distribution shows that the target population is widely accepting. But the big neutral category means that the things making learners' experiences different also need to be looked at.

Table 1.
Participants' Perceptions of ELSA Speak Effectiveness and Key Learning Outcomes (N = 110).

Variable	Category / Indicator	n (%)	M	SD
Perceived effectiveness of ELSA Speak	Effective / Very effective	85 (77.3%)	—	—
	Neutral	20 (18.2%)	—	—
	Ineffective	5 (4.5%)	—	—
Phoneme-level feedback usefulness	Detailed phoneme-level feedback	—	4.2	0.7
Real-time corrective feedback	Strongly agree / Agree	80 (72.7%)	—	—
Gamification and learning motivation	Reported increased motivation	75 (68.2%)	—	—
Confidence in pronunciation	Reported increased confidence	90 (81.8%)	—	—

Note: Percentages are calculated based on N = 110. M = mean; SD = standard deviation. Frequency counts (n) are reported only when percentages were provided by the study data.

Participants highly appreciated the software because it could give them detailed feedback on phonemes (M = 4.2, SD = 0.7) in a short time. In fact, 72.7% of them showed strong agreements or agreements that instant corrective feedback was the app's best feature. Also, 68.2% agreed that gamification elements like progress tracking, achievement badges, and daily challenges made them more motivated to learn. Confidence in pronunciation skills improved significantly, with 81.8% reporting enhanced self-assurance when speaking English after prolonged use of the application.

4.1.2. User Experience and Interface Design

Table 2.
Participants' Evaluation of ELSA Speak User Interface and Technical Issues (N = 110).

Variable	Category / Indicator	n (%)	M	SD
Overall user interface evaluation	User interface rating	—	3.8	0.8
Navigation design	Agreement on intuitive navigation	70 (63.6%)	—	—
Technical issues	Reported technical problems	30 (27.3%)	—	—
Speech recognition concerns	Inconsistencies with Vietnamese-accented English	—	—	—
Accuracy feedback concerns	Incorrect error flagging / failure to recognize improvement	—	—	—

Note: Percentages are based on N = 110. Frequency counts (n) are reported where percentages were provided. M = mean; SD = standard deviation. Qualitative indicators describe commonly reported user experiences without numerical quantification.

The majority of respondents who used the app said they preferred the user-friendly interface of the application (M = 3.8, SD = 0.8). 63.6% of those surveyed had an agreement on the easy-to-use navigation design and the visually appealing graphics that kept them interested. Nevertheless 27.3% claimed they had technical issues, especially with speech recognition which is not working well when processing Vietnamese-accented English. The individuals mentioned instances in which the system mistakenly identified proper pronunciation as incorrect pronunciations or failed to acknowledge improvements, which made them angry and confused about what the standards for accuracy were.

4.1.3. Challenges and Limitations

Table 3.
Reported Challenges and Limitations of ELSA Speak Use (N = 110).

Challenge / Limitation	Description	n (%)
Lack of contextualized practice	Exercises focused on isolated words or short phrases rather than connected speech in authentic contexts	45 (40.9%)
Cost of premium features	Subscription cost created accessibility barriers for some learners	40 (36.4%)
Internet connectivity requirements	Unstable network access affected application use	35 (31.8%)

Note: Percentages are calculated based on N = 110. Frequency counts (n) are reported from the provided percentages. Descriptions summarize participants' reported concerns.

Even though most of the respondents had positive views on the software, they also expressed its significant shortcomings in the process of using the app. Notably, 40.9% of the surveyed pointed out apprehensions about the lack of adequate contextualized practice opportunities. They highlighted that most of the exercises used isolated words or short phrases instead of connected speech in real-life situations. This made it hard for the learners to apply what they learned to real-life situations. Also, 36.4% said that the cost of premium features was a barrier for them to get, especially for those from lower-income families. Internet connectivity requirements (31.8%) proved challenging for students to learn in places where the network was not always stable.

4.2. Qualitative Findings

Based on 12 open-ended questions with 6 criteria such as (1) specific pronunciation improvements that ELSA Speak supported with, (2) the most helpful features for learning, (3) the challenges that came up, (5) how ELSA Speak compared to traditional teaching methods, and (6) suggestions for making it better. The transcripts were analyzed in four principal themes: (1) personalized learning and autonomy, (2) immediate feedback as motivational tool, (3) accent bias and speech recognition limitations, and (4) need for complementary human instruction.

4.2.1. Personalized Learning and Autonomy

The respondents highly appreciated ELSA Speak's benefits since it adapts each individual's level of speaking skills and learning style. A particular person said:

I like the app as it detects my specific weaknesses in pronunciation. For example, I have trouble in pronouncing the "th" sound, and the app produces lessons just for me to practice. Moreover, I can practice at my own pace without feeling embarrassed about making mistakes or being pressured by my classmates. This autonomy learning allows me to concentrate on what I really need most.

This theme actually fits with Self-Determination Theory. In fact, the theory demonstrates the way ELSA Speak meets learners' needs for autonomy by giving them self-directed, personalized practice opportunities that are not available in traditional classrooms, in which the curriculum focuses on group needs instead of individual ones.

4.2.2. Immediate Feedback as Motivational Tool

Through the interviews, instant corrective feedback is considered as the most important feature of ELSA Speak. The respondents stressed that immediate evaluation sped up fixing mistakes and learning new skills. One individual said:

I make mistakes with my pronunciation in regular classes, but I do not find out until much later when the teacher has time to correct everyone. I get right away with ELSA. I can try again immediately to see if I've gotten better. I feel motivated in practicing because I can see that I'm making progress right away.

The gamification features including progress bars, achievement badges, and daily streaks, are particularly efficient as they make learners interested in doing the same pronunciation exercises over and over again, which they might have found boring otherwise.

4.2.3. Accent Bias and Speech Recognition Limitations

A significant theme that emerged from the interviews was the systematic inconsistencies in speech recognition, especially for Vietnamese-accented English. People said they were frustrated when the system refused to acknowledge pronunciation they thought was correct or did accept versions they knew were wrong. One person said:

ELSA says I'm wrong when I say a word correctly, which my teacher said I did. Sometimes I know I messed up, but ELSA gives me a good grade. Because of this inconsistency, I'm not sure what is really right. I wonder whether am I getting better or just learning how to fool the app?

Another finding that backs up the respondents' worries is AI speech recognition systems being biased against certain accents. In fact, software programs are trained mostly on data from native speakers, so it is not easy to correctly assess speech with a non-native accent. This could lead to incorrect feedback that could throw off pronunciation development.

4.2.4. Need for Complementary Human Instruction

Even though ELSA Speak had many benefits in the process of pronunciation learning, the students always made sure that the app could not fully replace real teachers. All of the learners highly appreciated the way their teachers gave them context-based explanations, cultural insights, and opportunities to practice speaking that the app cannot provide. One person said:

ELSA is great for practicing certain sounds, but my teacher helps me understand why some pronunciations are important in real conversations. She tells me when to use formal and informal speech, how intonation can change the meaning of a word, and how well I communicate overall, not just how well I say each sound.

This theme emphasizes the importance of blended learning methods. When there is a support from the app together with real teachers, this definitely enables learners to take advantage of the unique benefits of each type of instruction.

5. Discussion

From the results of this study, ELSA Speak is proved as an effective tool for Vietnamese ELL learners to improve their pronunciation. Furthermore, the research simultaneously identifies major limitations and proposes the ways to enhancement. The mixture of quantitative and qualitative data are used to enhance the reliability of these conclusions, as both analysis methods identify consistent patterns for the app's benefits and challenges.

5.1. The Accent Bias Paradox

The findings of this study indicate strong positive perceptions, with 77.3% evaluating ELSA Speak as effective or very effective in accordance with previous research [8, 9]. However, positive perceptions do not automatically indicate pedagogical effectiveness. A careful investigation brings up an important point: the recipients said they were very happy with their learning, but they also had trouble understanding speech, especially Vietnamese-accented English.

Furthermore, this contradiction shows that there are some fundamental shortcomings in AI speech recognition technology. Bajorek [10] and Koenecke, et al. [11] demonstrate that speech recognition algorithms have a natural bias against accents, so they lead to reduced accuracy when processing speech with a non-native accent. Their effects on education could be important. First, insufficient recognition may result in false negatives, which means rejecting correct pronunciation as incorrect. This can make students less sure of themselves. Second, students may change how they say things to make the algorithm happy rather than making sure they are understood. Third, accent bias may adversely affect certain learners and create a feedback mechanism in which those facing difficulties receive less accurate guidance.

5.2. The Risk of Over-Reliance on Automated Feedback

72.7% of participants said that immediate feedback was the best thing about ELSA Speak. While instant corrective feedback is consistent with evidence-based principles of effective instruction, machine-readable feedback possesses fundamental pedagogical limitations that uncritical implementation fails to acknowledge. Automated feedback is based on rules and does not take context into consideration. It cannot take into account the context of communication, the expectations of the interlocutor, or the pragmatic factors that determine how well pronunciation works in real-life situations.

Also, automated feedback can hinder the growth of cognitive abilities. When the learners only rely on external algorithmic evaluation, they might not learn the way to monitor themselves, which is important for real-world communication where no app can fix mistakes right away. To learn how to pronounce words correctly, it is necessary to develop self-regulatory competence, which means being able to judge one's own pronunciation, find mistakes on one's own, and fix them [41]. If students depend too much on automated feedback, they may become learned helplessness and not be able to do anything without help from technology.

5.3. The Segmental-Suprasegmental Gap

A significant limitation insufficiently acknowledged in participants' favorable assessments relates to ELSA Speak's emphasis on segmental features (individual phonemes) to detrimental effect of suprasegmental features (stress, rhythm, intonation). Research studies consistently show that suprasegmental features are more important for comprehension than segmental accuracy [14, 15]. ELSA Speak claims that it operates on the suprasegmental features, but the way the app is set up focuses more on single words or short phrases than on connected speech in real conversations. Participants' requests for more contextualized practice scenarios (27.3% in survey data) indicate an implicit acknowledgment of this limitation.

6. Pedagogical Implications

These results have significant implications for educators and curriculum developers thinking about AI-driven pronunciation tools. First, blended learning should be seen as a need, not a choice. ELSA Speak should be seen as an extra tool for blended learning, not a replacement for teachers. Classroom time should focus on what AI cannot do such as contextualized communicative practice, suprasegmental feature development, cognitive-based instruction, and personalized feedback that takes into account each learner's goals and situations.

In addition, educators should teach students about the limits of technology in a clear way and accent bias and the restrictions of algorithms to help them become more knowledgeable in the digital world. It is essential that learners comprehend that errors in speech recognition may indicate system constraints rather than deficiencies in their pronunciation, thereby safeguarding against unjustified diminishment of confidence. Also, students ought to understand that just because they follow the rules does not imply they are good at communicating.

Third, classroom instruction should help students learn how to control themselves. Instruction need to clearly explain to students how to be aware of their own thinking and how to keep track of their own progress so they avoid depending too much on automated feedback. Technology should help these self-regulatory skills instead of replacing them. Fourth, designing the curriculum should put comprehension ahead of sounding like a native speaker. This pedagogical perspective, corroborated by recent pronunciation research [15] recognizes that intelligible, confidence-enhancing pronunciation better meets learners' communicative requirements than the pursuit of unobtainable and pedagogically challenging native-like standards.

7. Limitations

This study's limitations encompass the single-institution sample, which may restrict the generalizability to a wider Vietnamese EFL learner population. The limited qualitative sample (n = 24) constrains the depth of contextual comprehension, and the cross-sectional design inhibits longitudinal evaluation of enduring effects. The dependence on self-reported perceptions instead of objective pronunciation evaluation constitutes another limitation; however, this methodological choice was theoretically appropriate and facilitated a critical analysis of the gap between learner satisfaction and pedagogical value.

8. Conclusion

With the use of a simultaneous mixed-method design, this study investigated the perceptions of first-year English majors at Hanoi Metropolitan University towards implementing ELSA Speak for improving their pronunciation. The findings indicated that ELSA Speak is an effective tool for Vietnamese EFL learners to improve their pronunciation, especially with segmental problems because Vietnamese and English have different phonological rules. Moreover, most of the respondents had very positive opinions, with 77.3% of them saying the application was effective or very effective. Real-time feedback, interactive exercises, and gamification elements are the app’s most prominent features. 81.8% of learners felt much more confident and 68.2% found motivated with these features.

However, the respondents found a few major issues that needed to be fixed. For example, AI systems cannot always understand speech with regional accents, so in some situations AI systems are biased against certain accents. There were also problems with the interface and not enough practice opportunities that were relevant to the situation. It is the segmental-suprasegmental gap. Although effectively facilitating individual sound practice, ELSA Speak may insufficiently address suprasegmental features and authentic communicative contexts that are crucial for comprehensive pronunciation competence.

This study presents numerous practical implications for stakeholders in Vietnamese EFL education setting. Teachers should take advantage of ELSA Speak’s strengths in personalized, intensive phoneme practice while keeping the traditional role of teaching in developing communicative competence. Structured ELSA Speak practice into formal language programs could be integrated in providing explicit guidance on optimal usage strategies and clear pedagogical explanations that prioritize intelligibility-focused objectives over native-like standards. Application developers should also focus on improving speech recognition accuracy for different Vietnamese accents, making the interface easier to use, and adding more contextualized practice scenarios to make teaching more effective.

To get around these limitations, further research should be conducted in multi-institutional studies with larger, more diverse samples; longitudinal designs to examine lasting effects. Especially, learner perceptions with objective pronunciation metrics is essential to be combined in thorough assessments to evaluate both segmental and suprasegmental dimensions of pronunciation development. Additionally, It would also be helpful to compare ELSA Speak to other AI-based pronunciation apps to find the best technology for teaching. It is also important to analyze strategies for reducing accent bias in speech recognition systems and how explicit instruction affects AI's limitations on how students use technology and how their pronunciation changes.

References

[1] V. Braun and V. Clarke, "Using thematic analysis in psychology," *Qualitative Research in Psychology*, vol. 3, no. 2, pp. 77-101, 2006. https://doi.org/10.1191/1478088706qp063oa

[2] D. T. Tran, "Common English pronunciation problems among Vietnamese learners: A case of non-English majors," *Multidisciplinary Reviews*, vol. 7, no. 9, p. e2024224, 2024. https://doi.org/10.31893/multirev.2024227

[3] T. Do Anh, "Intelligible pronunciation: Teaching English to Vietnamese learners," *VNU Journal of Foreign Studies*, vol. 37, no. 1, 2021. https://doi.org/10.25073/2525-2445/vnufs.4666

[4] H. Van Van, "Interpreting MOET'S 2018 general education English curriculum," *VNU Journal of Foreign Studies*, vol. 38, no. 5, pp. 1-22, 2022. https://doi.org/10.25073/2525-2445/vnufs.4866

[5] J. Burston, "Twenty years of MALL project implementation: A meta-analysis of learning outcomes," *ReCALL*, vol. 27, no. 1, pp. 4-20, 2015. https://doi.org/10.1017/S0958344014000159

[6] A. Kukulska-Hulme, "Will mobile learning change language learning?," *ReCALL*, vol. 21, no. 2, pp. 157-165, 2009. https://doi.org/10.1017/S0958344009000202

[7] G. Stockwell, "Using mobile phones for vocabulary activities: Examining the effect of platform," *Language Learning & Technology*, vol. 14, no. 2, pp. 95–110, 2010.

[8] W. Asteriana, "The effectiveness of using ELSA speak application to teach English speaking skills to high school students," *Journal of English Education Forum*, vol. 5, no. 4, pp. 219–224, 2025. https://doi.org/10.29303/jeef.v5i4.927

[9] E. EKİNCİ and M. EKİNCİ, "Students learning English as a foreign language's views on mobile applications: a case study," *International Journal of Language Academy*, vol. 5, no. 18, pp. 175-193, 2024. https://doi.org/10.18033/ijla.3659

[10] J. P. Bajorek, "Voice recognition still has significant race and gender biases," Harvard Business Review, 2019. https://hbr.org/2019/05/voice-recognition-still-has-significant-race-and-gender-biases

[11] A. Koenecke *et al.*, "Racial disparities in automated speech recognition," *Proceedings of the National Academy of Sciences*, vol. 117, no. 14, pp. 7684-7689, 2020. https://doi.org/10.1073/pnas.1915768117

[12] F. D. Davis, "Perceived usefulness, perceived ease of use, and user acceptance of information technology," *MIS Quarterly*, vol. 13, no. 3, pp. 319-340, 1989. https://doi.org/10.2307/249008

[13] E. L. Deci and R. M. Ryan, "The" what" and" why" of goal pursuits: Human needs and the self-determination of behavior," *Psychological Inquiry*, vol. 11, no. 4, pp. 227-268, 2000. https://doi.org/10.1207/S15327965PLI1104_01

[14] T. M. Derwing and M. J. Munro, *Pronunciation fundamentals: Evidence-based perspectives for L2 teaching and research*. Amsterdam, Netherlands: John Benjamins, 2015.

[15] J. M. Levis, *Intelligibility, oral communication, and the teaching of pronunciation*. Cambridge, U.K: Cambridge University Press, 2018.

[16] P. M. Rogerson-Revell, "Computer-assisted pronunciation training (CAPT): Current issues and future directions," *Relc Journal*, vol. 52, no. 1, pp. 189-205, 2021.

[17] A. Dou, W. Xu, X. Li, S. Zhang, and J. Zhang, "Artificial intelligence in language learning: Bridging gaps, revealing patterns, and charting the future," *International Journal of Distance Education Technologies*, vol. 23, no. 1, pp. 1-24, 2025. https://doi.org/10.4018/IJDET.385045

[18] F. Yang, K. Li, and R. Li, "AI in language education: Enhancing learners' speaking awareness through AI-supported training," *International Journal of Information and Education Technology*, vol. 14, no. 6, pp. 828-833, 2024.

[19] J. Khlaisang and P. Sukavatee, "Mobile-assisted language learning to support English language communication among higher education learners in Thailand," *Electronic Journal of E-Learning*, vol. 21, no. 3, pp. 234-247, 2023. https://doi.org/10.34190/ejel.21.3.2974

[20] C. C. Lin and V. Lin, "Effects of MALL on L2 listening comprehension: A meta-analysis," *ReCALL*, vol. 31, no. 1, pp. 4–17, 2019.

[21] A. Bandura, *Social learning theory* (Englewood Cliffs, NJ, USA). Prentice Hall, 1977.

[22] J. Piaget, *The psychology of intelligence*. Totowa, NJ, USA: Littlefield Adams, 1976.

[23] S. Brown and R. Ahmed, "Adaptive learning algorithms in a mobile application for foreign language learning," *Information Systems and Networks*, vol. 18, no. 2, pp. 131–138, 2025. https://doi.org/10.23939/sisn2025.18.2.129

[24] G. Stockwell, *Technology and motivation in English-language teaching and learning. In International perspectives on motivation: Language learning and professional challenges*. London: Springer, 2013.

[25] Y. Zhao and C. Lai, "Technology and second language learning: Promises and problems," in *Technology-mediated learning environments for young English learners*: Routledge, 2023, pp. 167-206.

[26] I. S. Samad and I. Ismail, "ELSA speak application as a supporting media in enhancing students' pronunciation skill," *Majesty Journal*, vol. 2, no. 2, pp. 1-7, 2020.

[27] N. Fauziah, Z. Z. Aororah, R. A. Ramadhany, and S. M. Hamid, "The impact of the english language speech assistant (ELSA) application on students' pronunciation skills," *INTERACTION: Jurnal Pendidikan Bahasa*, vol. 11, no. 2, pp. 193-203, 2024. https://doi.org/10.36232/interactionjournal.v11i2.26

[28] S. R. Rahman, N. Nurhamdah, and M. Munawir, "ELSA Speak Application to improve the students' pronunciation at member of libam," *Al-Irsyad: Journal of Education Science*, vol. 3, no. 1, pp. 15-21, 2024.

[29] A. P. Wilujing and D. Karikawati, "The use of ELSA speak application to improve pronunciation accuracy of ninth graders at SMPN 1 Gondang," *Semantik: Jurnal Riset Ilmu Pendidikan, Bahasa dan Budaya*, vol. 3, no. 2, pp. 340–347, 2025.

[30] E. Elsani, R. Salsabila, M. F. I. Putra, N. K. Nabila, and D. Nahartini, "The effect of using ELSA Speak app for first-semester students' English speaking proficiency," *Edukatif: Jurnal Ilmu Pendidikan*, vol. 5, no. 6, pp. 2644-2655, 2023.

[31] S. A. Karim, A. Q. S. Hamzah, N. M. Anjani, J. Prianti, and I. G. Sihole, "Promoting EFL studentsâ€™ speaking performance through ELSA speak: An artificial intelligence in english language learning," *Journal of Languages and Language Teaching*, vol. 11, no. 4, pp. 655-668, 2023. https://doi.org/10.33394/jollt.v11i4.8958

[32] D. Salsabilla and E. Rosmiyati, "Teaching speaking by using'elsa ai'to the eighth grade students of smp negeri 42 palembang," *Esteem Journal of English Education Study Programme*, vol. 7, no. 2, pp. 841-849, 2024.

[33] E. Horwitz, "Language anxiety and achievement," *Annual Review of Applied Linguistics*, vol. 21, pp. 112-126, 2001. https://doi.org/10.1017/S0267190501000071

[34] A. Mardiah and S. Saadillah, "Maximizing ELSA speak for developing english fluency and reducing speaking barriers in language learners," *Issues In Applied Linguistics & Language Teaching* vol. 7, no. 1, pp. 262-271, 2025.

[35] F. Saragih and S. Lubis, "Students' language anxiety levels and factors in using ELSA Speak for grade eight at SMP Santo Thomas 3 Medan," *TRANSFORM Journal of English Language Teaching and Learning*, vol. 14, no. 2, pp. 116–124, 2025.

[36] ELSA Speak, "ELSA speak: AI-powered English pronunciation coach," 2023. https://elsaspeak.com

[37] T. N. Pham and T. X. Dang, "An investigation into the application of artificial intelligence for language teaching and learning in Vietnam," *Journal of Contemporary Educational Policies and Practices*, vol. 9, no. 2, pp. 265-283, 2025. https://doi.org/10.52296/vje.2025.557nb

[38] N. T. Tran, T. T. Nguyen, and P. H. Tran, "Transforming foreign language acquisition: A perception-based study on artificial intelligence in language education across Asia," *International Journal of TESOL Studies*, vol. 8, no. 2, pp. 6-28, 2026. https://doi.org/10.58304/ijts.250914

[39] J. W. Creswell, *Research design: Qualitative, quantitative, and mixed methods approaches*, 4th ed. Thousand Oaks, CA, USA: Sage, 2014.

[40] K. Malterud, V. D. Siersma, and A. D. Guassora, "Sample size in qualitative interview studies: guided by information power," *Qualitative Health Research*, vol. 26, no. 13, pp. 1753-1760, 2016.

[41] K. Saito and L. Plonsky, "Effects of second language pronunciation teaching revisited: A proposed measurement framework and meta-analysis," *Language Learning*, vol. 69, no. 3, pp. 652-708, 2019. https://doi.org/10.1111/lang.12345