Volume 8 Number 3 (2026)
September-December 2026
Page: 302-320

E-ISSN: 2656-4491
P-ISSN: 2656-4548
DOI:10.37680/scaffolding.v8i3.10884

SCAFFOLDING
Jurnal Pendidikan Islam dan Multikulturalisme

USING GEMINI VOICE-TO-TEXT IN FACILITATING PRONUNCIATION AWARENESS THROUGH SELF-CORRECTION AMONG EIGHTH-GRADE STUDENTS

Mutiara Ni'mah[1], Widyastuti[2], Ahmad Munir[3]
[123]Universitas Negeri Surabaya; Indonesia
Correspondence E-mail; mutiaranimah285@gmail.com

Submitted: 29/09/2026 Revised: 16/10/2026 Accepted: 24/10/2026 Published: 29/10/2026

Abstract

This study aims to investigate how Gemini Voice-to-Text facilitates pronunciation awareness through self-correction among eighth-grade students in an English as a Foreign Language (EFL) classroom context. Specifically, it examines students' experiences using the tool, how AI-generated transcription supports their awareness of pronunciation differences, and the practical benefits and challenges experienced during implementation. This study used a qualitative case study design. The participants in the study were eight eighth-grade students from SMP Negeri 16 Surabaya, East Java, Indonesia. These students participated in pronunciation activities using Gemini Voice-to-Text. Data were collected through classroom observation, semi-structured interviews, and documentation of the AI-generated transcription output. Analysis of the collected data was conducted thematically with data familiarization, coding, categorization, and developing themes. The findings revealed that students used Gemini Voice-to-Text through three interconnected processes: accessing the tool during speaking activities, reviewing AI-generated transcription, and comparing their spoken production with the recognized output. These processes provided students with opportunities to notice pronunciation differences and engage in self-correction through repetition and modification of their pronunciation. Students also perceived the tool as beneficial because it provided immediate feedback, supported independent practice, and increased opportunities for repeated pronunciation practice. On the other hand, there were some problems with using the tool, especially regarding mistakes in the transcription due to varying ways people say words, how fast someone speaks, and the amount of noise in the classroom. This research concludes that Gemini Voice-to-Text can facilitate pronunciation awareness through self-correction when its feedback is critically interpreted by students and supported by teacher guidance.

Keywords Artificial Intelligence, EFL Learning, Gemini Voice-to-Text, Pronunciation Awareness, Self-correction.

© 2026 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution-NonCommercial 4.0 International License (CC BY NC) license (https://creativecommons.org/licenses/by-nc/4.0/).

INTRODUCTION

Pronunciation plays an important role in intelligible communication because it affects how effectively learners convey their intended meanings. In contemporary times, pronunciation instruction has shifted its focus from mimicking how native speakers speak to providing students with the ability to communicate clearly and be understood by others who may not share the same first language (Galante & Piccardo, 2022; Loewen & Sato, 2024). Therefore, it also affects the intelligibility and communicative effectiveness of learners. It is especially important for learners to keep track of their own pronunciation, assess it, and correct their mistakes during speaking(Saito, 2019). Nevertheless, EFL learners have difficulty with this as they do not possess the ability to recognize the differences between their produced speech and the target pronunciation. Additionally, classroom assistance provided by teachers is limited due to time constraints during lessons and, therefore, does not allow for individualized monitoring.

The researcher found this problem during preliminary observations of students of SMP Negeri 16 Surabaya. Eight eighth-grade students stated that they lacked confidence in speaking English because they were worried about making mistakes with their pronunciation. In accordance with the classroom data presented in the original article, approximately 75% of the students from each of the six eighth-grade classes had spoken language performance results that were higher than the average score for the class. However, we need to be careful with how we interpret these results; they do not provide proof that fear of mispronouncing words led to poorer spoken language performance. Fear of mispronouncing words could prevent students from speaking completely; this would limit the amount of practice they receive to enhance their skills. Research has indicated that speaking anxiety is a significant variable affecting second language oral production (Mora et al., 2024). It is likely that creating an environment where students can try out their pronunciation, evaluate their own spoken products, and correct themselves without having to worry about being judged by others would help improve their speaking ability. Therefore, a technology that allows students to assess their pronunciation privately appears to be a viable option.

Automatic Speech Recognition (ASR) converts spoken language into a visual representation and offers instant feedback (Ngo et al., 2024). A meta-analysis of 15 studies reported an overall positive effect of ASR on pronunciation performance, although the outcomes varied according to feedback type, instructional duration, and learner characteristics. Likewise, in their systematic review of EFL pronunciation research, Liu et al. (2025) found that the majority of current studies on

this topic focus on the accuracy of pronunciation, and often use a quasi-experimental design. A study of 75 Japanese college students, utilizing two 25-minute practice periods using speech-to-text software, demonstrated that the percentage of correct transcriptions increased from 38% to 62%, and students felt more confident in speaking (Henry, 2025). Similarly, Indonesian college students utilizing Gemini AI for speaking practice also indicated that they felt comfortable practicing without concern for negative social evaluations (Nurilahiah et al., 2026). However, Widodo et al. (2026) found that a participant who struggled to operate the voice-interaction feature achieved only a 1% increase in speaking scores, compared with an average increase of 13% among participants overall. These findings suggest that ASR can provide opportunities for pronunciation practice and feedback, but its benefits depend partly on learners' ability to use and interpret the technology.

A study with secondary school students reported that using a structured self-evaluation process helped students change how they diagnosed their speaking ability, and develop improvement plans (De Vrind et al., 2024). If there is a mismatch between what a student intended to say and what is shown on the transcript, then it cannot be assumed that the student has made a pronunciation error. There are other factors, such as the student's accent, the clarity of his/her speech, and any background noise, that could affect the accuracy of the transcription (Emara & Shaker, 2024; Tong et al., 2022).

Research comparing automatic speech recognition with human judgments of second-language intelligibility also indicates the importance of distinguishing machine recognition from human understanding (Inceoglu et al., 2023). Furthermore, Gemini Voice-to-Text is a general-purpose AI-supported speech input feature rather than a dedicated phonetic assessment instrument. Therefore, students will have to decide for themselves if the discrepancy was caused by their own speech or by the tool. It remains unknown how young EFL learners make these decisions, particularly in Indonesian junior high school settings.

The second was a mixed-methods explanatory sequential design study conducted to find that ASR, when used in conjunction with peer-correction, can enhance both L2 speaking abilities and L2 pronunciation for EFL students (Sun, 2023). However, what was missing from this study was the use of self-correction through the student's reading of the recognition output, rather than relying on peer-correction.

The third study demonstrated a 13% average increase in undergraduate speaking scores for those using Gemini AI within task-based learning. They also identified several limitations, including

interruption of the learning process, irrelevant responses from the AI, and incompatibility with certain devices(Widodo et al., 2026). The research gap is that while the study examined the speaking fluency of university students, it did not examine how the use of Gemini could develop awareness of pronunciation.

Fourth, the researcher conducted interviews with six Indonesian college students and found that the students had positive perceptions of Gemini AI for pronunciation, fluency, vocabulary, and confidence; however, the students identified two areas for improvement — the instability of the Internet connection and the "artificial" sounding nature of the response (Nurilahiah et al., 2026). What is missing from this research is an observation of how the students were using the tool, rather than simply asking them about their perception of speaking skills overall.

Fifth, Baskara (2025) Conducted a qualitative study examining the use of ChatGPT and Gemini by sophomore students from Indonesia. He identified several limitations of using both tools, such as poorly calibrated language difficulty, rigid conversational exchange, and a reliance on a stable internet connection. The limitation was that it investigated the overall effectiveness of these two chatbots as teaching resources, rather than focusing on their potential to improve pronunciation. Therefore, the current study seeks to narrow the scope of investigation to only one tool, Gemini Voice-to-Text, and one specific skill, pronunciation.

These studies reveal three gaps. The first is the studies of speech-to-text and Gemini, which had participants confirmed to be university students (Baskara, 2025; Henry, 2025; Nurilahiah et al., 2026; Widodo et al., 2026) left out junior high school students. The second deficiency is that the Gemini studies investigated the overall performance of speaking or perception, but did not use the transcripts as sources of pronunciation feedback. The third deficiency is that none of the five studies examined how learners identified errors in their own production and corrected them. Therefore, this study aims to investigate how Gemini-voice-to-Text facilitates pronunciation awareness through self-correction among eighth-grade students, focusing on students' experiences, the influence of AI-generated feedback on pronunciation awareness development, and the practical benefits and challenges encountered during its implementation.

The theoretical basis of this study is rooted in both pronunciation instruction pedagogy and second language acquisition. It highlights the importance of enhancing the awareness of the learner's pronunciation features through meaningful practice and feedback (Celce-Murcia et al., 2010). Similarly, Levis (2018) in Intelligibility, Oral Communication, and the Teaching of Pronunciation,

identifies intelligibility as one of the primary goals of pronunciation instruction. The two viewpoints are supported by Schmidt (1990), who suggests that a person needs to pay attention to linguistic features when acquiring a language, and the self-regulation learning model (Zimmerman, 2002), which illustrates how learners observe and modify their performance. It helps to explain how Gemini Voice-to-Text can increase a student's awareness of pronunciation by allowing them to view what they say against what the AI transcribes, notice where there might be differences between what they intended to say and what was actually transcribed, and reflect on ways to correct themselves.

The purpose of this research is to examine the ways in which Gemini Voice-to-Text assists in developing awareness of pronunciation through self-correction by 8th-grade students at SMP Negeri 16 Surabaya, East Java, Indonesia. The research specifically focused on the experiences of the students when utilizing Gemini Voice-to-Text in pronunciation practice; the manner in which AI-generated transcripts help support the development of awareness of pronunciation and the subsequent attempts at self-correction; and the perceived benefits and challenges associated with the use of Gemini Voice-to-Text.

METHOD

This study used a qualitative case study to examine eighth-grade students' perspectives regarding their experiences with and perceptions of using Gemini Voice to Text for pronunciation practice at SMP Negeri 16 Surabaya in East Java, Indonesia. The eight students chosen to participate in the study were those who had taken part in and successfully finished the pronunciation exercises. Information collection involved participant observation, semi-structured in-depth interviews, and documentation of AI-generated transcripts. On August 14th, 2026, a classroom observation took place with emphasis placed upon students' interactions with Gemini, such as practicing speaking skills, checking transcripts, finding errors in transcription, and trying to fix those errors. In addition, interviews were utilized to elicit information concerning students' experiences, potential advantages, and possible disadvantages associated with Gemini Voice to Text; whereas documentations were used to provide examples of discrepancies between what students intended to say and what was transcribed by Gemini Voice to Text. To determine patterns of awareness of correct pronunciation and self-correction, thematic analysis was applied to the collected data. Analysis of the data involved the following steps: familiarization with the data, coding the data, categorizing the data, and developing themes from the categorized data. Data triangulation,

member checking, and an audit trail were employed as means of ensuring the trustworthiness of the study. Ethical considerations related to this study were based on voluntary participation, anonymity, and confidentiality.

FINDINGS AND DISCUSSION

Findings

Students' Experiences with Gemini Voice-to-Text in Pronunciation Activities.

The first theme illustrates the observation of students using Gemini Voice-to-text for pronunciation practice, supported by interviews and transcription documentation. The findings revealed three main activities: accessing the application, reviewing AI-generated transcriptions, and comparing spoken production with the transcription results. The observed behaviors and their interpretations are presented in Table 1.

Table 1. Students' Interaction with Gemini Voice-to-Text during Pronunciation Activities

No.	Observation Category	Observation Indicator	Observed Students' Behaviors	Interpretation of Students' Experiences
1	Interaction with Gemini Voice-to-Text	Students access Gemini Voice-to-Text during speaking activities.	Students independently prepared their devices, opened the Gemini application, and activated the microphone feature before practicing spoken English.	Students demonstrated their ability to use Gemini Voice-to-Text independently as a digital resource for pronunciation practice.
2	Interaction with Gemini Voice-to-Text	Students interact with AI-generated transcription results.	Students read the generated transcriptions, reviewed the displayed text, and focused on particular words or phrases after speaking.	Students used AI-generated transcription as immediate visual feedback to examine their spoken production.
3	Interaction with Gemini Voice-to-Text	Students compare spoken output with transcription results.	Students compared intended words with transcriptions, identified mismatches, repeated unclear words, and discussed discrepancies with	Students engaged in self-monitoring and attempted self-correction by evaluating differences between intended speech and AI-generated transcription.

	peers or the teacher.

Source: Adopted from Kohnke et al. (2023), Zhang and Meifang (2024)

During the observation (Observation Datum 1), students prepared their device independently, opened the Gemini application, and activated the microphone feature to input voice and produce spoken English. However, the interaction illustrates that students were able to integrate Gemini into their pronunciation practice independently without relying fully on teacher support. In this way, students got a taste of Gemini as an accessible digital resource to be directly applied in a speaking activity.

This time, students not only read the text on the screen but also looked out for keywords that appeared differently from the intended pronunciation. The transcription results served as prompt feedback that enabled students to evaluate their spoken performance. Data from the students' interviews supported the findings of Datum 2. Informant 4 explained that Gemini Voice-to-Text enabled students to see the words and sentences they had pronounced in written form. Similarly, Informant 8 Datum 2 stated the transcription results as a visual representation of their spoken English. These responses reflect the perception by students that Gemini served as a tool that made their pronunciation visible through AI-generated text with which they could interrogate their spoken production more transparently.

The third learning experience derived from the observation marker of "students compare their oral output with transcriptions." While observing, students compared their own words with the text generated by Gemini, noted differences in spelling and pronunciation of challenging words as they reread them out loud regularly, and discussed variations with each other or teachers. Students compared their spoken production with the transcription output. It was also evident in the way students repeated their unclear pronunciation when transcribing items that did not produce the desired sounds. Observation data suggested that students assessed whether the mismatch was due to their pronunciation, or, in this case, to AI recognition limits of the transcript, affected by classroom noise.

Interview data supported this finding. Gemini was also referred to by Informant 6 as a way to check if pronunciation errors were made, and it enabled comparison between transcription results and original sentences before determining if correction was required. The experience of Informant 6 Datum 3 showed that students referred to the output produced by Gemini and then performed self-correction. Differences between students' spoken production and Gemini recognition results supported this observation in the transcription documentation. Several pronunciation mismatches

were found, such as Informant 2 Datum 3 described an instance in which the intended word *sheep* was transcribed as *ship*, while Informant 8 Datum 3 reported a mismatch involving *three* and *tree*.

Such mismatches often became observable evidence for students to catch discrepancies between their intended pronunciation and AI recognition. This experience was supported further by the findings of the interviews conducted. As Informant 1 said, having a tool like Gemini made it easier to point out errors in pronunciation since the transcription results would show when words were spoken but failed to match their expected recognition. Datum 3 Informant 1 appears to see AI-generated transcription as feedback. Similarly, Informant 4 Datum 3 explained that differences between students' intended words and AI-generated transcriptions helped them become aware of potential pronunciation problems. Furthermore, Informant 7 Datum 3 explained that Gemini enabled students to practice speaking without waiting for teacher feedback on every word. It indicated that for some students, Gemini worked quite literally as a tool for independent pronunciation practice bolstered by rapid feedback.

2. Fostering Pronunciation Awareness Through Self-Correction with Gemini Voice-To-Text

The second theme describes the way that Gemini Voice-To-Text provided students with feedback on their awareness of pronunciation using an AI-generated transcription. This was based upon both the interviews and observations. It was found that there were three distinct ways that students developed a better understanding of what might be difficult about pronouncing certain words or phrases. Those included:

1. Reorganization of phonetic process errors due to wrong transcription output;
2. Detecting mismatches between spoken phrases and transcription produced by Gemini
3. Reflecting on pronunciation accuracy after receiving AI feedback.

Table 2. Students' Pronunciation Awareness through Self-Correction Using Gemini Voice-to-Text

No.	Interview Category	Interview Indicator	Students' Reported Experiences
1	Recognizing potential pronunciation errors	Students recognize possible pronunciation difficulties after reviewing transcription feedback.	Students reported becoming aware of words or sentences that might have been pronounced unclearly when Gemini generated unexpected transcriptions.
2	Identifying transcription mismatches	Students compare intended words with Gemini-generated transcription results.	Students identified differences involving words such as think–thing, ship–sheep, school–cool, and tree–three.
3	Reflecting on	Students repeat and	Students reported repeating

	pronunciation through self-correction	adjust their pronunciation after reviewing transcription feedback.	words or sentences, adjusting articulation, stress, and intonation, and checking subsequent transcription results.

Source: Adopted from Wang and Hamill (2021), Yan et al. (2024), Zhang and Meifang (2024)

The interview findings demonstrated that students were now paying closer attention to words and sentences that they felt may not have been clearly spoken. Informant 1 (Interview 1, Datum 1) stated that the transcriptions allowed him/her to see which words/sentences needed more practice in terms of pronunciation. Additionally, Informant 3 (Interview 1, Datum 1) identified that poor pronunciation can result in Gemini producing words/phrases that were unexpected. Lastly, Informant 6 (Interview 1, Datum 1) realized there could be pronunciation problems that he/she had not noticed before. Therefore, it appears that the use of AI-generated transcriptions gave students the ability to reflect upon what they produced orally and pinpoint pronunciation areas that required additional attention. It should be noted, however, that the results reflect the students' views regarding pronunciation problems and do not provide evidence of an objective analysis of the students' actual pronunciation errors.

The second part focused on students' ability to identify when their intended spoken words did not match Gemini's transcription of those same words. The interviews provided numerous examples of students encountering transcription errors while practicing pronunciation. Informant 2 (Interview 1, Datum 2) stated that the intended words think and ship were frequently transcribed as thing and sheep. Also, Informant 3 (Interview 1, Datum 2) stated that an unexpected transcription of think was due to poor articulation. Informant 4 (Interview 1, Datum 2) explained a mismatch where school was transcribed as cool. Informant 8 (Interview 1, Datum 2) indicated that the tree was sometimes recognized as three. Additionally, documentation of transcription outputs provided additional support for the interview findings by showing us that there were inconsistencies between students' intended speech and Gemini's recognition. However, these inconsistencies need to be viewed with caution as they could be caused by either the student having difficulty with pronunciation or the speech recognition system having limitations.

In the third part, students tried to adjust how they pronounced words based on the feedback from the artificial intelligence program. According to interviews, students reacted to unexpected transcriptions by repeating words or phrases and changing how they said them. For example, Informant 1 (Interview 1, Datum 3) stated that he practiced each word/phrase over and over again

until Gemini understood what he wanted to say. Informant 2 (Interview 1, Datum 3) also reported that he altered the way he stressed and used intonation when he repeated the same phrase, which demonstrates that he was focusing on suprasegmental elements of speech in addition to the actual sounds. Likewise, Informant 7 (Interview 1, Datum 3) continued practicing his pronunciation because he felt it would help him perform better in speaking. Nevertheless, just because Gemini was able to recognize a student's pronunciation does not necessarily mean that the student improved their pronunciation skills.

Before using the app, Informant 1 (interview 1, datum 3) stated that she was not aware of the pronunciation issues she faced; however, she became much more aware of potential mistakes once she received the instant transcription feedback. It is evident from this example that the opportunity for students to see the transcription of their speech allowed them to be more involved in assessing their own spoken output. In addition to being corrected by teachers, students were able to view the transcription of their speech, compare it to what they intended to say, and determine whether they needed to do more practice on their pronunciation. Overall, these results show that Gemini Voice-to-Text created opportunities for self-monitoring and student-initiated correction during pronunciation activities in the classroom.

3. Students' Perceptions on the Practical Benefits and Limitations of AI-Supported Pronunciation Tool

The third theme consisted of described benefits and challenges experienced by eighth graders in using Gemini Voice-to-Text to facilitate pronunciation awareness through self-correction. Data from the interviews with eight students identified Gemini Voice-to-Text as a suitable tool to support them in learning pronunciation, through providing instant feedback, making it easier to identify problems with their pronunciation, enabling them to hear their recordings repeated, and learn by themselves. Yet, at the same time, students named difficulties related to the accuracy of speech recognition, interpretation of AI feedback, and some settings that had an impact on transcript results. These perceived benefits and limitations are summarized in Table 3.

Table 3. Students' Perceptions on the Practical Benefits and Limitations of an AI-Supported Pronunciation Tool

No.	Interview Category	Interview Indicator	Students' Reported Experiences
1	Practical Benefits	Immediate transcription feedback and pronunciation awareness	Students reported using displayed transcriptions to identify possible pronunciation

			difficulties and compare recognized words with their intended speech.
2	Practical Benefits	Independent pronunciation practice	Students reported practicing spoken English individually without always requiring immediate teacher correction.
3	Practical Benefits	Repeated practice and confidence	Students reported repeating words after receiving transcription feedback and feeling more confident when practicing independently.
4	Practical Limitations	Speech recognition accuracy	Students reported unexpected transcription results when their speech was unclear, too fast, or not immediately recognized.
5	Practical Limitations	Interpretation of AI-generated feedback	Students experienced uncertainty about whether transcription mismatches resulted from pronunciation difficulties or system recognition errors.
6	Practical Limitations	Environmental noise	Students reported that background noise interfered with speech recognition and produced unexpected transcription results.

Source: Adopted from Shadiev (2020), Kohnke et al. (2023)

Practical Benefits of fostering pronunciation awareness through self-correction with Gemini Voice-to-Text

Informant 1 (Interview 2, Datum 1) stated that they used the transcription provided by Gemini as a baseline for determining that their spoken production was accurate. Informant 4 (Interview 2, Datum 2) realized that even though they had previously thought he/she was producing certain words accurately, those words needed additional practice. Informant 8 (Interview 2, Datum 3) also noted that comparing their spoken production with the transcription results allowed them to be more aware of the differences in pronunciation between what they produced and what Gemini transcribed. However, although the transcription results were based on how Gemini perceived the spoken input, they do not provide a definitive measure of the student's pronunciation accuracy.

Interview data showed that poor articulation, excessive speed of speech, and delay in recognizing the student's speech were factors contributing to inconsistent transcription results. Informant 1 (Interview 2, Datum 6) stated that at times, Gemini transcribed unintended words due to unclear speech. Informant 3 (Interview 2, Datum 7) described speaking too fast as a cause of

transcription errors, whereas Informant 6 (Interview 2, Datum 7) experienced delays in recognition of what they were saying. Overall, the above comments illustrate that students had inconsistent speech recognition experiences during pronunciation practice sessions.

Informant 3 (Interview 2, Datum 7) was unsure if the reason why the transcription did not match the student's pronunciation was due to poor pronunciation on the part of the student or if Gemini had failed to accurately transcribe the student's speech. Similarly, Informant 6 (Interview 2, Datum 7) stated that while Gemini helped locate possible areas where she may have had difficulty with her pronunciation, the transcribing results produced by Gemini were not always correct. Finally, Informant 8 (Interview 2, Datum 7) emphasized how crucial it is for teachers to provide additional explanation to students when they are unsure as to why their pronunciation was recognized differently than expected via transcription feedback. Therefore, teacher assistance continues to be necessary in order for students to be able to assess feedback, distinguish between potential pronunciation issues and recognition errors, and ultimately choose appropriate remedial actions.

In addition to environmental noise being a real-world constraint for using Gemini Voice-to-Text during pronunciation exercises, it was also found to be a problem. Informant 2 (Interview 2, Datum 8) stated that the noise level at times affected his ability to transcribe the text. On the other hand, Informant 7 (Interview 2, Datum 8) stated that environmental noise contributed to making errors in recognizing words correctly. If environmental noise interferes with the accuracy of speech recognition, students may obtain unexpected outcomes when attempting to clearly articulate specific words. Therefore, environmental factors should be taken into account when evaluating transcription feedback and designing pronunciation exercises utilizing AI-based tools.

Discussions

The first research question investigated how eighth-grade students' experiences with the use of Gemini Voice-to-Text during pronunciation activities. The results showed that Gemini functioned as an AI-assisted feedback tool that enabled students to observe, evaluate, and reflect on their spoken production. By interacting with this transcription output, students compare their intended pronunciation with the words detected by the system. These experiences illustrated that Gemini Voice-to-Text was seen not only as a speech transcription tool but also as a platform that offered students the space to analyze their pronunciation in speaking exercises.

This finding is consistent with previous research emphasizing the potential of automatic speech recognition (ASR) to provide accessible feedback during second-language pronunciation practice. The ASR-based tools provide learners with instant visual cues of spoken language, comparing their pronunciation versus that produced by a system-generated output (Ngo et al., 2024). The results from previous research on the use of automatic speech recognition (ASR) to support pronunciation learning indicate that providing students with written transcriptions of what they say provides the student with individualized practice and feedback on their spoken language (Evers & Chen, 2022; Yaniafari et al., 2022). Previous studies have also investigated the outcome of mobile-assisted pronunciation feedback as well as the underlying interactional processes that lead to the reported outcomes(Dai & Wu, 2022). It is consistent with the findings of this study. For the purposes of this study, Gemini Voice-to-Text performed a similar function by transcribing students' oral production into written form. This was necessary as spoken language is ephemeral, and often learners find it hard to analyze the language material after production.

Also, the view of learner autonomy and self-monitoring can be considered. The results indicated that students were engaged in rating their own pronunciation and not solely relying on the teacher's corrections. When students listened to Gemini transcription results, they performed a watch-and-decide process where they made decisions about whether or not their pronunciation needed more practice. This reading aligns with self-regulated learning theory, which characterizes learners as those who observe their own learning patterns, appraise their knowledge and behaviors during a learning task, and manage the execution of the evaluated behaviors/metacognitive processes (Ebedy, 2025; Panadero, 2017).

As highlighted in the literature review, learners were likely to continue repeating pronunciation difficulties when opportunities for immediate feedback were limited (Leow, 2023; Saito, 2019). The use of AI tools that include voice input is said to provide an opportunity to overcome the issue of rapid feedback through the ability to give individual feedback on the student's response (Du & Daniel, 2024; Huang, 2024; Zheng & Yang, 2024). Gemini Voice-to-Text was used for additional tasks if students needed a quick feedback loop during speaking activities.

Furthermore, the results also suggest that students' experience involved more than simply interacting with the technology. They needed to interpret what the transcription results meant by reflecting on whether the differences were due to the pronunciation produced by students or limitations of the recognition system. It focuses on this finding as ASR systems can be affected by

variation in pronunciation, accent, and environmental conditions, which include classroom noise (Emara & Shaker, 2024; Tong et al., 2022). In addition, learners have identified potential benefits and challenges associated with AI-based speech evaluation for speaking practice (Zou et al., 2024). Therefore, the experience for students was not just getting AI feedback, but critically evaluating what that feedback meant.

The second research question explored the impact of using Gemini Voice-to-Text on eighth-grade students' awareness of pronunciation through self-correction. The results indicated that students noticed pronunciation differences between their spoken production and the transcription output provided by Gemini. Zhang (2022) interpreted the process in relation to his Noticing Hypothesis that learners need to pay attention to some aspect of language before they can learn. The learner cannot influence the parts of language that go unnoticed. For this study, Gemini transcription served as an external noticing trigger in that it represented students' spoken output in writing. This representation allowed students to recognize some differences they might not otherwise have.

From a metacognitive perspective, learners (a) need to recognize the problem; (b) typically evaluate their performance or product based on existing evidence and decide whether to keep working or not; and (c) may adjust their work (Zimmerman, 2002). Furthermore, the results suggested that pronunciation awareness was related to students' self-correction attempts. Noticing pronunciation differences, students reported having repeated words, changing pronunciation, and trying other forms of saying the same word. This is indicative of the noticing and self-correcting process.

This aligns with feedback theory, which argues that good feedback enables learners to identify the gap between their current performance level and the desired level of performance. (Carless & Winstone, 2023). Therefore, Gemini Voice-to-Text is not an automatic correction tool in this process, but a basis of feedback that supports students in their reflection processes. The transcription output offered students some insight they could use to check on how their pronunciation was coming along.

The results should be viewed with caution, however. Just because transcription differences exist, it does not imply that pronunciation was recognized incorrectly as errors by an ASR system. Errors can also happen due to accent change, diversity in pronunciations, or extrinsic factors. If pronunciation varies, it may affect ASR recognition because the system relies upon acoustic features

of speech input, which is suggested by previous studies (Tong et al., 2022). However, while Gemini Voice-to-Text provided stages for awareness of pronunciation through opportunities to notice and reflect upon assigned speech, students are not passive, and teacher support is necessary at all levels of this process.

The third research question asked about the advantages and disadvantages of students trying to self-correct their pronunciation using Gemini Voice-to-Text. The results revealed several advantages for the students who used it, such as immediate feedback, recognizing difficulties in pronunciation, and practicing independently.

Another advantage was the fact that we could repeat for pronunciation practice. They were able to imitate words after hearing differences in pronunciation, students reported. This experience is an example of the principles of self-regulated learning, in which learners use feedback to self-monitor and adapt their learning (Panadero, 2017). Students engaged in a process of repeated cycles of speaking, reviewing their recordings, and re-speaking. Research on the use of interactive AI applications shows that they may increase pronunciation accuracy and that these increases can be sustained (Xodabande et al., 2025). Although the current study did not measure accuracy, the repeated speak-review-respeak cycle described by students is similar to the cycles that have been shown to lead to improved performance.

Bashori et al. (2024) stated that AI-supported language environments may provide learners with less threatening opportunities to practice speaking and receive feedback. This is consistent with the finding that students built confidence while practicing pronunciation since they could receive AI feedback without the pressure of being judged immediately by classmates or teachers. Fear of making mistakes with regard to pronunciation is a common reason why students experience speaking anxiety in junior high school (Marpaung & Fithriani, 2023). Practice with AI-mediated tools has been positively correlated with increased willingness to communicate and low anxiety speaking practice for young learners (Fathi et al., 2024; Tai & Chen, 2024).

However, the students also faced many challenges in addition to these benefits. The biggest problem is transcribing accuracy. Students also said that Gemini occasionally translated to a word different from what they meant to say. This mismatch was most apparent in cases where pronunciation differences, speaking speed, or noise from the environment influenced recognition results. This result is consistent with previous work showing that ASR-based pronunciation tools have certain technological limitations. Speech recognition accuracy is influenced by background

noise, variation in pronunciation, and speech clarity (Emara & Shaker, 2024). Consequently, although ASR systems can be used to provide feedback opportunities, the transcription of this output can still not necessarily be considered an accurate reflection of learners' pronunciation.

Another challenge was that students had to have enough interpretation skills to know whether the problem was with their pronunciation or with the AI system. This demonstrates how important mediation by the teacher is in learning, strongly supported by AI. Feedback provided by AI can only give us some extra info, but learners still need the help of instructors to make sense of the feedback and correct themselves. The findings are consistent with the idea that AI-based learning tools, such as ChatGPT-type tools, should be used in conjunction with human-provided instruction (Mompean, 2024) and that users need to assess critically the output generated by AI (Kasneci et al., 2023). The technology facilitated the initiation of pronunciation feedback through increased accessibility, and effective learning was reliant on students actively working with the data and critically interpreting their own needs, as well as the support they received from teachers.

CONCLUSION

The present study investigated eighth-grade students' experience of developing pronunciation awareness through using Gemini Voice-to-Text. The students described Gemini Voice-to-Text as an AI-supported feedback tool that enabled them to see their spoken production through AI-transcribed text. This enabled student to contrast what they intended to say with what was transcribed, to identify pronunciation challenges and errors, and to reflect on their spoken performance in lessons. The results also suggested that students went through noticing, self-monitoring, and self-correcting processes when using Gemini Voice-to-Text. The students reported that the transcription feedback enabled them to identify pronunciation problems, repeat words that were not clear, and practice their pronunciation independently. However, students faced challenges as well, such as the accuracy of transcription, variation in pronunciation, speed of speaking, and noise in class. In conclusion, Gemini Voice-to-Text can be considered an AI-supported learning resource that provides students with opportunities to receive pronunciation feedback, observe their spoken production, and engage in reflective self-correction. In reality, as a tool to help with student practice on their own while receiving more feedback during the learning task, Gemini better fits as a supplement than replacing teacher feedback.

REFERENCES

Bashori, M., van Hout, R., Strik, H., & Cucchiarini, C. (2024). I Can Speak: Improving English Pronunciation Through Automatic Speech Recognition-Based Language Learning Systems. *Innovation in Language Learning and Teaching*, *18*(5), 443-461. https://doi.org/10.1080/17501229.2024.2315101

Baskara, F. X. R. (2025). ChatGPT and Google Gemini in EFL Education: A Qualitative Exploration of Pedagogical Efficacy among Indonesian Sophomores. *JOLLT Journal of Languages and Language Teaching*, *13*(1), 436-447. https://doi.org/10.33394/jollt.v13i1.9926

Carless, D., & Winstone, N. (2023). Teacher Feedback Literacy and its Interplay with Student Feedback Literacy. *Teaching in Higher Education*, *28*(1), 150-163. https://doi.org/10.1080/13562517.2020.1782372

Celce-Murcia, M., Brinton, D. M., & Goodwin, J. M. (2010). *Teaching Pronunciation Hardback with Audio CDs (2): A Course Book and Reference Guide*. Cambridge University Press.

Dai, Y., & Wu, Z. (2022). Mobile-Assisted Peer Feedback on EFL Pronunciation: Outcome Effects, Interactional Processes, and Shaping Factors. *System*, *111*, 102953. https://doi.org/10.1016/j.system.2022.102953

De Vrind, E., Janssen, F. J. J. M., Van Driel, J. H., & De Jong, N. H. (2024). Improving Self-Regulated Learning of Speaking Skills in Foreign Languages. *The Modern Language Journal*, *108*(3), 601-624. https://doi.org/https://doi.org/10.1111/modl.12953

Du, J., & Daniel, B. K. (2024). Transforming Language Education: A Systematic Review of AI-Powered Chatbots for English as a Foreign Language Speaking Practice. *Computers and Education: Artificial Intelligence*, *6*, 100230. https://doi.org/https://doi.org/10.1016/j.caeai.2024.100230

Ebedy, H. G. M. (2025). The Effect of Self-Monitoring Strategy in Developing Pronunciation Among EFL Majors: Links to self-esteem. *Bsu-Journal of Pedagogy and Curriculum*, *4*(7), 1-25. https://doi.org/10.21608/bsujpc.2025.351432.1062

Emara, I., & Shaker, N. (2024). The impact of non-native English Speakers' Phonological and Prosodic Features on Automatic Speech Recognition Accuracy. *Speech Communication*. https://doi.org/10.1016/j.specom.2024.103038

Evers, K., & Chen, S. (2022). Effects of an Automatic Speech Recognition System with Peer Feedback on Pronunciation Instruction for Adults. *Computer Assisted Language Learning*, *35*, 1869-1889.

Fathi, J., Rahimi, M., & Derakhshan, A. (2024). Improving EFL Learners' Speaking Skills and Willingness to Communicate Via Artificial Intelligence-Mediated Interactions. *System*, *121*, 103254. https://doi.org/https://doi.org/10.1016/j.system.2024.103254

Galante, A., & Piccardo, E. (2022). Teaching Pronunciation: Toward Intelligibility and comprehensibility. *ELT Journal*, *76*(3), 375-386. https://doi.org/10.1093/elt/ccab060

Henry, J. (2025). Speech-To-Text Technology as a Tool For Pronunciation Assessment and Self-Correction in EFL Contexts: A Pilot Study. *Elite: English and Literature Journal*, *12*(1), 68-79.

Huang, J. (2024). Enhancing EFL Speaking Feedback with ChatGPT's Voice Prompts. *International Journal of TESOL Studies*, *6*(3), 4-23. https://doi.org/10.58304/ijts.20240302

Inceoglu, S., Chen, W.-H., & Lim, H. (2023). Assessment of L2 intelligibility: Comparing L1 listeners and Automatic Speech Recognition. *Recall*, *35*(1), 89-104. https://doi.org/10.1017/S0958344022000192

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F.,…Kasneci, G. (2023). ChatGPT for good? On opportunities and Challenges of Large Language Models for Education. *Learning and Individual Differences*, *103*, 102274.

https://doi.org/10.1016/j.lindif.2023.102274

Kohnke, L., Moorhouse, B. L., & Zou, D. (2023). ChatGPT for Language Teaching and Learning. *Relc Journal*, *54*(2), 537-550. https://doi.org/10.1177/00336882231162868

Leow, R. P. (2023). Noticing hypothesis. *The TESOL Encyclopedia of English Language Teaching. https://doi. org/10.1002/9781118784235. eelt0086. pub2*.

Levis, J. M. (2018). *Intelligibility, Oral Communication, and the Teaching of Pronunciation*. Cambridge University Press.

Liu, Y., binti Ab Rahman, F., & binti Mohamad Zain, F. (2025). A Systematic Literature Review of Research on Automatic Speech Recognition in EFL Pronunciation. *Cogent Education*, *12*(1), 2466288.

Loewen, S., & Sato, M. (2024). Pronunciation. In S. Loewen & M. Sato (Eds.), *A Practical Guide to Second Language Teaching and Learning* (pp. 142-162). Cambridge University Press. https://doi.org/DOI: 10.1017/9781108780520.008

Marpaung, N. R., & Fithriani, R. (2023). English Speaking Anxiety Among Indonesian Junior High School Students: Unveiling Causes and Solutions. *Tell-us Journal*, *9*(1), 146-163. https://doi.org/10.22202/tus.2023.v9i1.6678

Mompean, J. A. (2024). ChatGPT for L2 Pronunciation Teaching and Learning. *ELT Journal*. https://doi.org/10.1093/elt/ccae050

Mora, J. C., Mora-Plaza, I., & Bermejo Miranda, G. (2024). Speaking Anxiety and Task Complexity Effects on Second Language Speech. *International Journal of Applied Linguistics*, *34*(1), 292-315. https://doi.org/https://doi.org/10.1111/ijal.12494

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.

Nurilahiah, S., Hasim, W., & Dewi, S. (2026). Exploring Indonesian EFL Students' Perceptions of Gemini AI in Developing English Speaking Skills. *Pendas: Jurnal Ilmiah Pendidikan Dasar*, *11*, 279-286. https://doi.org/ https://doi.org/10.23969/jp.v11i02.48598

Panadero, E. (2017). A Review of Self-regulated Learning: Six Models and Four Directions for Research [Review]. *Frontiers in Psychology, Volume 8 - 2017*. https://doi.org/10.3389/fpsyg.2017.00422

Saito, K. (2019). Individual Differences in Second Language Speech Learning in Classroom Settings: Roles of Awareness in the Longitudinal Development of Japanese Learners' English /ɹ/ pronunciation. *Second Language Research*, *35*(2), 149-172. https://doi.org/10.1177/0267658318768342

Schmidt, R. W. (1990). The Role of Consciousness in Second Language Learning1. *Applied Linguistics*, *11*(2), 129-158. https://doi.org/10.1093/applin/11.2.129

Shadiev, R., & Yang, M. (2020). Review of Studies on Technology-Enhanced Language Learning and Teaching. *Sustainable Education and Approaches*, *12*(2), 524. https://doi.org/https://doi.org/10.3390/su12020524

Sun, W. (2023). The Impact of Automatic Speech Recognition Technology on Second Language Pronunciation and Speaking Skills of EFL Learners: A Mixed Methods Investigation. *Frontiers in Psychology*, *14*. https://doi.org/10.3389/fpsyg.2023.1210187

Tai, T. Y., & Chen, H. H. J. (2024). Improving elementary EFL Speaking Skills with Generative AI Chatbots: Exploring Individual and Paired Interactions. *Computers & Education*, *220*, 105112. https://doi.org/10.1016/j.compedu.2024.105112

Tong, F., Li, T., Liao, D., Xia, S., Li, S., Hong, Q., & Li, L. (2022). The Xmuspeech System for Accented English Automatic Speech Recognition. *12*(3). https://doi.org/https://doi.org/10.3390/app12031478

Wang, X., & Hamill, C. (2021, 13-17 Dec. 2021). Automatic Generation of Diagnostic Content Feedback in Spoken Language Learning and Assessment. 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU),

Widodo, A., Vivianti, V., Karim, S. A., & Polim, H. (2026). Task-Based Learning Reimagined: How Gemini AI Enhances EFL Students' Speaking Fluency in Self-Directed Learning. *JOLLT Journal of Languages and Language Teaching*, *14*(2), 596-611. https://doi.org/10.33394/jollt.v14i2.18207

Xodabande, I., Shiri, S., & Zohrabi, M. (2025). Exploring the Impacts of an AI-driven Instructional Intervention on Iranian EFL Learners' Pronunciation Skill Development. *Discover Education*, *4*, 307. https://doi.org/10.1007/s44217-025-00782-2

Yan, H., Singh, M. K. S., & Rawian, R. M. (2024). AI-Based Corrective Feedback in EFL Interactive Speaking: Insights from Interactionist SLA Theory. *Educational Sciences: Theory&Practice, 24(No. 1 (2024): Educational Sciences: Theory and Practice*. https://doi.org/https://doi.org/10.12738/

Yaniafari, R. P., Olivia, V., & Suharyadi. (2022). The potential of ASR for improving English pronunciation: A review. *KnE Social Sciences*.

Zhang, & Meifang. (2024). Enhancing Self-Regulation and Learner Engagement in L2 Speaking: Exploring the Potential of Intelligent Personal Assistants Within a Learning-Oriented Feedback Framework. *BMC Psychology*, *12*(421). https://doi.org/10.1186/s40359-024-01917-0

Zhang, Z. (2022). Noticing in Second Language Acquisition. *BCP Education & Psychology*, *7*, 184-190. https://doi.org/10.54691/bcpep.v7i.2634

Zheng, L., & Yang, Y. (2024). Research Perspectives and Trends in Artificial Intelligence-Enhanced Language Education: A review. *Heliyon*, *10*(19), e38617. https://doi.org/10.1016/j.heliyon.2024.e38617

Zimmerman, B. (2002). Becoming a Self-Regulated Learner: An Overview. *Theory Into Practice*, *41*, 64-70. https://doi.org/10.1207/s15430421tip4102_2

Zou, B., Wang, C., Yan, Y., Du, X., & Ji, Y. (2024). Exploring English as a Foreign Language Learners' Adoption and Utilisation of ChatGPT for Speaking Practice Through an Extended Technology Acceptance Model. *International Journal of Applied Linguistics*, *35*(2), 689-704. https://doi.org/10.1111/ijal.12658