Linguistic Research 43(2): 563-592
DOI: 10.17250/khisli.43.2.202606.009

Korean L2 learners' English vowel adjustments during interactions with two ASR systems: An AI assistant and a speech-to-text tool*

Liying Bai** · **Jae Yung Song***
(Chung-Ang University)

Bai, Liying and Jae Yung Song. 2026. Korean L2 learners' English vowel adjustments during interactions with two ASR systems: An AI assistant and a speech-to-text tool. *Linguistic Research* 43(2): 563-592. There has been growing interest in using Automatic Speech Recognition (ASR) systems to support second-language (L2) pronunciation practice. This study examined how Korean learners of English adjusted their pronunciation in response to two ASR feedback modalities: auditory feedback from an AI assistant (Amazon Alexa) and visual text feedback from a speech-to-text tool (Google Docs). Nineteen Korean L2 learners produced 36 words containing six English vowels (/i, ɪ, ɛ, æ, ɑ, ʌ/). Following a misrecognition, they repeated the same word up to two additional times. Results from mixed-effects models showed that (1) recognition rates generally increased across attempts for most vowels, with the largest gains occurring between the first and second attempts; (2) when modifying their pronunciation, learners primarily relied on vowel lengthening rather than spectral adjustments; and (3) the two ASR systems appeared to elicit somewhat different adjustment patterns. Overall, ASR provided effective real-time feedback that prompted learners to modify their pronunciation. However, in the absence of explicit guidance, these adjustments were largely confined to temporal changes. These findings highlight the need for ASR activities and feedback to be carefully designed to facilitate pronunciation learning. **(Chung-Ang University)**

Keywords ASR feedback, vowel pronunciation, phonetic adjustment, Korean EFL learners, L2 English

* This research was supported by the Chung-Ang University Research Scholarship Grants in 2024.
** First Author
*** Corresponding author

1. Introduction

Pronunciation plays a central role in successful communication, shaping how clearly second-language (L2) speakers are understood and how effectively they can convey their intended meaning to speakers from other cultural backgrounds (Couper 2021; Levis 2022). Yet, traditional English as a Foreign Language (EFL) classroom environments often face practical challenges, such as large class sizes and limited instructional time. These constraints make it difficult to provide sufficient, timely, and individualized feedback on pronunciation (Cucchiarini et al. 2012; Bashori et al. 2024). Because individualized feedback is critical for pronunciation improvement, the lack of it can slow learners' progress and may lead to communication breakdowns and increased anxiety (Baran-Łucarz and Lee 2021; Tsang 2025).

Automatic Speech Recognition (ASR) technology has emerged as a highly promising solution to these pedagogical challenges (Dizon 2020; Evers and Chen 2022; Bashori et al. 2024; Yang, Lai, and Chen 2024). By automatically transcribing learners' speech, ASR creates an autonomous learning environment where learners can receive instant feedback, monitor their progress, and practice as often as needed. These opportunities are difficult to achieve in large or time-constrained classes (McCrocklin 2016; Liakin et al. 2017; Inceoglu 2022). In addition to supporting practice, ASR-based tools can help diagnose pronunciation problems by identifying areas that need improvement. In a mobile-assisted pronunciation study that included an ASR-only condition (i.e., dictation-based ASR feedback without teacher feedback), learners still improved over time in comprehensibility, segmental accuracy, and word-stress accuracy, suggesting that ASR feedback can play a meaningful role in pronunciation learning when instructor correction is not available (Dai and Wu 2023). Although ASR systems are not entirely free from recognition errors (Cucchiarini and Strik 2017), learners generally perceive ASR-based pronunciation practice positively. For example, it can reduce speaking anxiety and increase confidence by freeing learners from teacher evaluation pressure and reducing fear of making mistakes (Dizon 2020; Bashori et al. 2024).

Building on this potential, a growing body of research has explored how ASR systems can support L2 learning. These efforts include vocabulary and pronunciation training through ASR-equipped websites (Bashori et al. 2024), gamified language-learning platforms such as Duolingo (Shortt et al. 2023), as well as more

general-purpose speech recognition tools. Such tools include speech-to-text (STT) tools, like Google Docs, and AI personal assistants, including Amazon Alexa, Apple Siri, and conversational AI models like ChatGPT. These technologies differ in how learners interact with them, ranging from isolated word recognition to continuous conversational input. They also differ in the types of feedback they provide, offering learners distinct interactive experiences. In the present study, we focused on two of the ASR systems, an STT tool (Google Docs) and an AI personal assistant (Amazon Alexa), to examine how learners adjust their pronunciation when engaging with different types of ASR systems in controlled word-production tasks.

Many general-purpose document editors, such as Google Docs, include voice-input features that convert speech to text. With their user-friendly and accessible interfaces, these editors can serve as practical and effective tools for autonomous pronunciation practice. Learners' spoken utterances are converted into visible text on the screen in real-time (Gutz et al. 2023), allowing them to immediately see whether their pronunciation matches the target (Leis 2025). Unlike conventional STT tools, voice-activated AI personal assistants, such as Amazon Alexa, respond to spoken input and provide auditory and interactive feedback (Dizon 2023). Their conversational interfaces allow learners to receive feedback in a manner that closely resembles interaction with a human interlocutor, offering unique educational potential for L2 pronunciation learning (Moussalli and Cardoso 2021; Dizon 2023). This human-like interaction is reflected in learners' perceptions; studies have shown that learners often treat AI assistants as "interlocutors" or "social partners," which could in turn prompt a more natural conversational style (Purington et al. 2017).

When misrecognitions occur, AI assistants prompt learners to identify errors and adjust their pronunciation through repetition and rephrasing (Dizon 2017; Song et al. 2022). For example, Song et al. (2022) investigated interactions between Korean L2 English speakers and Amazon Alexa. The study showed that Alexa's recognition rates for English vowels (/i/, /ɪ/, /æ/, /ɛ/) produced by Korean L2 learners was significantly lower (55%) than for those produced by native speakers of English (98%). Notably, when Alexa misrecognized a word, L2 speakers adjusted their subsequent pronunciations by modifying vowel duration, F1, and F2 in directions that enhanced the vowels' phonological features (for example, producing higher F2 for the front vowel /i/ compared with the initial misrecognition).

Together, these two forms of ASR systems offer unique opportunities to explore

their potential for supporting L2 pronunciation practice. As noted earlier, a key distinction between these systems lies in their feedback modalities. Alexa provides auditory, interactive feedback through verbal responses, while Google Docs STT offers visual, text-based feedback through text conversion. As a result, Google Docs tend to elicit more careful pronunciation by allowing learners to monitor their speech visually through text. In contrast, AI personal assistants, such as Amazon Alexa, facilitate interactive, socially oriented practice by providing auditory feedback and prompting repetition and rephrasing within more natural communicative exchanges.

Taking advantage of these different strengths, this study investigated how learners modify their pronunciations when interacting with the two ASR systems. In particular, it addressed three main research questions. First, does L2 learners' word production accuracy improve over the course of multiple pronunciation attempts following feedback from the two ASR systems? Second, when recognition fails, do learners modify the acoustic properties of their vowels (duration, F1, and F2) across subsequent attempts in ways that differ from their initial misrecognized production? Third, how do differences between the two ASR systems influence L2 learners' pronunciation adjustments?

By systematically addressing these questions, this study aims to provide deeper insights into how L2 learners adapt their speech in response to ASR-based feedback in autonomous, teacher-free learning environments. The findings could also offer practical implications for designing more effective and personalized ASR-based pronunciation activities for EFL learners.

2. Methods

2.1 Procedure

We initially recruited 20 participants (6 male, 14 female), all native speakers of Korean learning English as a second language. Data from one participant were excluded due to failure to follow the experimental procedure, resulting in a final dataset of 19 participants. Before the experiment, participants completed a brief questionnaire designed to collect basic demographic and language background information. Participants' ages ranged from 20 to 31 years (M = 23.8, SD = 3.31). All participants

were enrolled in undergraduate or graduate programs at universities in Seoul, with 12 undergraduates, 6 master's students, and 1 doctoral student. The participants' English proficiency was self-rated using the Common European Framework of Reference for Languages (CEFR). The distribution was A1-beginner (n = 1), A2-pre-intermediate (n = 8), B1-intermediate (n = 7), and B2-upper- intermediate (n = 3), with no C-level participants (C1-advanced/C2-proficient = 0). When the CEFR levels were converted to a numerical scale (A1 =1, A2 = 2, B1 = 3, B2 = 4, C1 = 5, C2 = 6), the mean proficiency score was 2.63, corresponding approximately to the A2-B1 range (lower-intermediate range). All participants had learned English in an EFL context. None reported living in English-speaking countries, except for two participants who had short-term stays (3 and 6 months).

2.2 Stimuli

The target stimuli consisted of 36 monosyllabic CVC words: 6 vowels (/i/, /ɪ/, /ɛ/, /æ/, /ɑ/, /ʌ/) × 6 words. One-third of these words formed minimal pairs for /i/ and /ɪ/, such as *beat* /bit/ and *bit* /bɪt/. Another one-third of the words formed minimal pairs for /ɛ/ and /æ/, such as *bet* /bɛt/ and *bat* /bæt/, while the remaining one-third formed minimal pairs for /ɑ/ and /ʌ/, such as *cop* /kɑp/ and *cup* /kʌp/. These three vowel contrasts were selected because they not only represent well-documented areas of difficulty for Korean learners of English, but also allowed for the construction of a sufficient number of minimal pairs that met the phonological and lexical constraints of the experimental design. The /i–ɪ/ and /ɛ–æ/ contrasts are not phonemically distinguished in Korean (Flege et al. 1997). The /ɑ–ʌ/ contrast, although present in the Korean vowel inventory, is acoustically distinct from its closest Korean counterparts (Yang 1996). The filler words, which were not included in the analyses, consisted of 12 monosyllabic CVC words: 3 vowels (/eɪ/, /aɪ/, /oʊ/) × 4 words (see Appendix for a complete list of stimuli). These diphthongs were selected because they are generally less problematic for Korean learners and were therefore expected to elicit fewer pronunciation errors, thereby reducing potential delays and keeping the experimental procedure efficient.

We chose stimuli ending in voiceless stops (/p/, /t/, or /k/). Our pilot data indicated that speakers often partially devoiced word-final voiced stops, leading Alexa to

mistakenly identify them as voiceless. For this reason, voiced stops were avoided in the word-final position. We also avoided the approximants /j/, /w/, /l/, and /ɹ/ because it is generally difficult to separate them from adjacent vowels (Harrington 2010). Lexical frequency was also taken into account in selecting the stimuli. Following Miller, Raney, and Demos (2020), who defined high-frequency words as those occurring 60 or more times per million, 21 of the 36 target words met this criterion based on frequency estimates from the Corpus of Contemporary American English (COCA). Eight words (*bat, bet, cop, mess, peak, pat, pet, seek*) occurred between 15 and 60 times per million, and the remaining seven (*dip, dock, duck, gut, hut, mat, nut*) occurred between 6 and 15 times per million. Although all stimuli were simple CVC words and were expected to be familiar to the participants, familiarity was verified prior to the experiment. Participants reviewed the word list and indicated any unfamiliar items. When participants reported an unfamiliar word, which happened only infrequently, they could consult dictionaries or online resources, with additional clarification provided by the researcher.

2.3 Procedure

The experiment consisted of two parts. One part used an Alexa device, Amazon's AI personal assistant, while the other part used Google Docs with its STT tool. The order of the two parts was counterbalanced across participants. Each part lasted approximately 20 minutes, with a five-minute break between parts. Both the Alexa device and Google Docs operated with American English settings to ensure consistency in speech recognition across systems. Because the two systems differ in their primary interaction formats, the elicitation procedure in each part was designed to reflect their typical use. As a voice-activated conversational assistant, Alexa requires a wake word and command structure; therefore, participants produced each target word within a sentence frame (e.g., "Echo, spell ___"). In contrast, in the Google Docs STT condition, producing words within sentences could introduce contextual influences, as STT systems often draw on surrounding linguistic context to generate transcriptions. To minimize these contextual effects and ensure that recognition was based primarily on the acoustic realization of the target word, participants produced each target word in isolation. The procedures for each part are described in more detail below.

2.3.1 Part 1: Alexa

We used an Alexa device (second-generation Echo Show) to assess participants' pronunciation. Participants were given a printed list of the stimuli in randomized order. For each word on the list, they said, "Echo, spell _____," prompting Alexa to spell the word. *Echo* served as the device's wake word. To minimize coarticulatory effects, participants were instructed to insert a brief pause between "Echo, spell" and the target word.

For instance, if a participant correctly pronounced the word *pat* in "Echo, spell *pat*," Alexa would respond, "Pat is spelled P-A-T." However, if the participant produced the vowel /æ/ in *pat* inaccurately, for example, by substituting /ɛ/, Alexa might respond, "Pet is spelled P-E-T." This type of substitution is common among Korean learners of English and often results in misrecognition of the intended word. When a word was recognized correctly, the participant immediately proceeded to the next word. Thus, no additional attempts were made following a successful first attempt, and no third attempt was made if recognition was successful on the second attempt. A third attempt was allowed only when both the first and second attempts were unsuccessful. If the word was still not recognized on the third attempt, the participant proceeded to the next word without further repetitions. Both the participants' speech and Alexa's responses were recorded in a sound-attenuated booth using a Logitech Blue Yeti USB microphone connected to a lab computer. The recordings were made in Audacity at a sampling rate of 48 kHz.

2.3.2 Part 2: Google Docs

In Part 2, participants produced the same set of words as in Part 1, but Google Docs was used in place of Amazon Alexa. After activating the STT tool of Google Docs on the lab computer, participants read the 48 words one at a time, referring to the same pre-printed word list used previously. For each word, participants activated the STT tool by clicking the microphone icon, produced the target word, and then clicked the button again to deactivate recognition before moving on to the next item. This reset ensured that each token was treated as an independent input. Without resetting the STT tool, the transcription engine would continue processing running speech and might generate output based on surrounding context or on its expectations about

what word should come next, rather than on the learner's actual pronunciation of the target word.

As participants pronounced the words, Google Docs' STT function automatically transcribed them and displayed the recognized words on the screen. For instance, if a participant correctly pronounced *pat*, Google Docs displayed *pat* as text on the screen. However, if the vowel /æ/ in *pat* was mispronounced as /ɛ/, the STT tool might display *pet* instead, indicating misrecognition of the target word. When the target word was correctly recognized, the participant immediately proceeded to the next word. In the event of misrecognition, the participant was allowed to repeat the same word, with up to two additional attempts permitted. The same recording setup and equipment were used as in Part 1. In addition, screen activity was recorded using the built-in screen capture and recording utilities of the Windows operating system to document the STT transcriptions displayed during the experiment.

2.4 Coding

Each target word was produced up to three times. When a word was correctly recognized at an earlier attempt, subsequent attempts were omitted and the task proceeded to the next target word. Thus, the number of attempts per word varied between 1 and 3, resulting in a total of 2,668 attempts. Of these, 2,267 attempts were included in the final analyses (first attempt: 1,207; second attempt: 584; third attempt: 476). All included attempts were coded as either 'correct' or 'incorrect.' When the participant's intended pronunciation of the target word matched the response from Alexa or Google Docs, the attempt was coded as 'correct' (e.g., *pet* /pɛt/ → *pet* /pɛt/) (Alexa n = 389, Google Docs n = 392). Two types of mismatches were identified, both of which were coded as 'incorrect.' Importantly, both types involved deviations in the vowel from the intended target. The first involved vowel mispronunciation with correctly produced consonants (e.g., *pet* /pɛt/ → *pat* /pæt/) (Alexa n = 509, Google Docs n = 631). As expected, this type of misrecognition occurred most frequently, as the experimental stimuli were designed to elicit systematic variation in vowel production. The second type involved misrecognition of both the vowel and consonant (e.g., *pet* /pɛt/ → *pad* /pæd/) (Alexa n = 170, Google Docs n = 176).

In addition, a total of 401 attempts were excluded. These included cases in which consonants were mispronounced despite accurate vowel production (e.g., *pet* /pɛt/

→ *bet* /bɛt/) (Alexa n = 109, Google Docs n = 51), cases in which Alexa produced a system message indicating recognition failure (e.g., "Sorry, I don't know that one.") (Alexa n = 40), and other attempts of the same target word associated with these two cases (Alexa n = 136, Google Docs n = 48). Finally, 17 attempts (Alexa n = 4, Google Docs n = 13) were excluded due to technical problems, poor acoustic quality, or extra productions.

For the final dataset, participants' vowel productions were also acoustically coded. Vowel onset and offset were identified in Praat, and the interval between these points was taken as vowel duration. F1 and F2 were measured at the temporal midpoint of each vowel. For statistical analyses, vowel duration (in milliseconds) was log-transformed to reduce skewness and the influence of extreme values, and F1 and F2 values (in Hz) were Lobanov-normalized (z-score transformed) within speakers.

2.5 Statistical analysis

We conducted two analyses. Analysis 1 examined whether recognition accuracy improved across attempts in the two STT systems. Analysis 2 investigated learners' acoustic adjustments following initial misrecognition in the two STT systems. All statistical analyses were performed in R 4.2.3 (R Core Team 2023) using mixed-effects regression models. In both analyses, CONDITION (Alexa vs. Google Docs) and ATTEMPT (first vs. second vs. third) were included as fixed effects, with participants and items as random intercepts. Both categorical predictors were sum-coded. Initially, maximal random-effects structures were considered (Barr et al. 2013); however, due to the unbalanced data structure and the resulting convergence and singularity issues, the final models adopted a simplified random-effects structure (Matuschek et al. 2017).

In Analysis 1, vowel recognition accuracy (correct vs. incorrect) served as a binary dependent variable, and models were fitted using the *glmer* function from the *lme4* package (Bates et al. 2015). In Analysis 2, the dependent variables were three acoustic measures (vowel duration, F1, and F2) and models were fitted using the *lmer* function from the same package. Data were analyzed using separate mixed-effects models for each of the six vowels. This approach was particularly appropriate for Analysis 2, as it allowed us to capture the direction and magnitude of spectral changes (F1 and F2) for individual vowels differing in vowel height and advancement.

To determine the significance of fixed effects and their interactions, we employed

Type III ANOVAs as implemented in the *car* package (Fox and Weisberg 2018). Because ATTEMPT included three levels, any significant main effect of this factor, as well as any significant ATTEMPT × CONDITION interaction, was followed up by post-hoc pairwise comparisons using the *emmeans* function in the *emmeans* package (Lenth 2022). All *p*-values were Bonferroni-adjusted to correct for multiple comparisons.

3. Results

3.1 Analysis 1: Recognition accuracy

We first examined the effects of ATTEMPT and CONDITION on vowel recognition accuracy. To assess whether recognition improved relative to the first attempt, recognition accuracy was calculated cumulatively across attempts. For example, if 5 out of 10 words (50%) were correctly recognized on the first attempt and 2 additional words were recognized on the second attempt among those initially misrecognized, the cumulative recognition rate at the second attempt was calculated as 7 out of 10 (70%). This cumulative method was used to reflect the ultimate success rate over multiple attempts. When accuracy is calculated independently for each attempt, no change in subsequent attempts may be misinterpreted as a decline in accuracy, because previously recognized items are no longer included in the calculation.

Figure 1 compares the cumulative recognition accuracy of Alexa and Google Docs for each of the six vowels across the three attempts. Table 1 presents the corresponding mixed-effects regression results for the same vowels: (a) /i/, (b) /ɪ/, (c) /ɛ/, (d) /æ/, (e) /ɑ/, and (f) /ʌ/. The effect of ATTEMPT on recognition accuracy was significant for five of the six vowel categories (except for /ɪ/; see Table 1). Pairwise comparisons using the *emmeans* function revealed a consistent pattern across vowels. For all five vowels, both the second and third attempts were recognized significantly more accurately than the first (all $p < 0.05$). Notably, the most substantial increase in recognition accuracy occurred between the first and second attempts. In contrast, no significant differences were found between the second and third attempts for any vowel, suggesting stabilization after the second attempt. There was also a significant ATTEMPT × CONDITION interaction for /i/ and /ɛ/. Pairwise comparisons using

the *emmeans* function revealed that for these two vowels, recognition accuracy improved significantly from the first to the second attempt and from the first to the third attempt in the Google Docs condition, but not in the Alexa condition. The effect of CONDITION was also significant for three vowels (see Table 1). Across attempts, recognition accuracy was higher in the Alexa condition than in the Google Docs condition for /i/, /ɪ/, and /ʌ/ (see Figure 1).

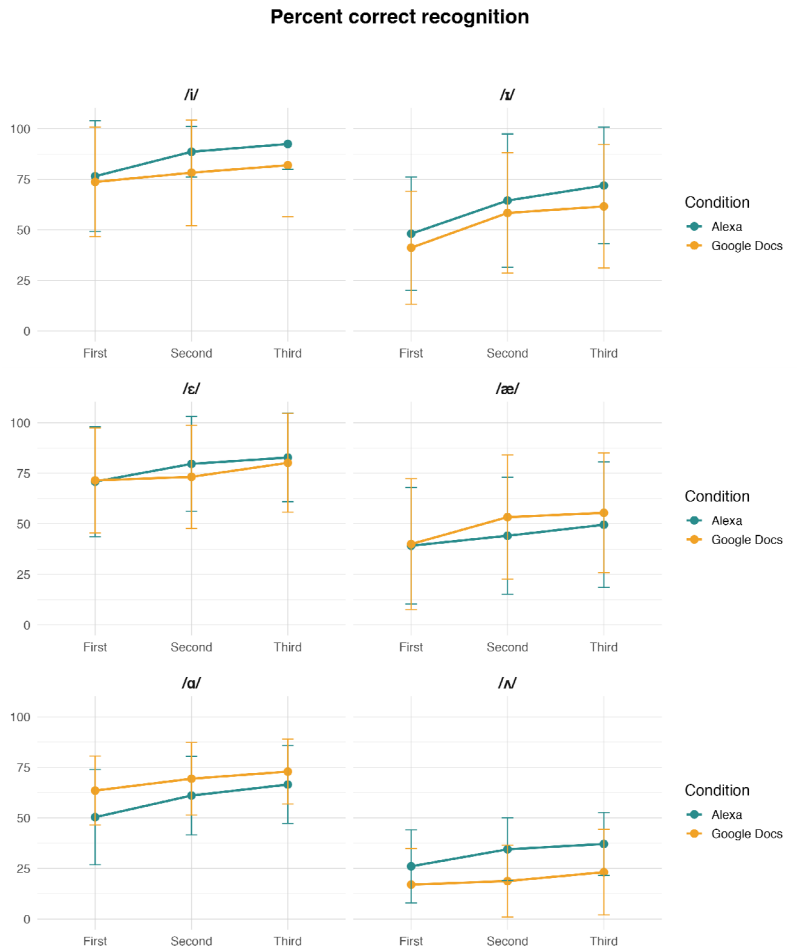

Figure 1. Percent correct recognition of the six vowels across attempts, separated by condition. Error bars represent standard deviations.

Table 1. Results of statistical analyses examining the effects of ATTEMPT and CONDITION on recognition accuracy

	(a) /i/			(b) /ɪ/		
	χ2	*df*	*p*-value	*χ2*	*df*	*p*-value
(Intercept)	1.62	1	0.20	0.74	1	0.39
Attempt	13.03	2	< 0.01	4.66	2	0.10
Condition	11.59	1	< 0.001	5.91	1	< 0.05
Attempt x Condition	7.61	2	< 0.05	2.38	2	0.30

	(c) /ɛ/			(d) /æ/		
	χ2	*df*	*p*-value	*χ2*	*df*	*p*-value
(Intercept)	0.56	1	0.46	14.31	1	< 0.001
Attempt	33.38	2	< 0.001	26.07	2	< 0.001
Condition	2.47	1	0.12	0.02	1	0.89
Attempt x Condition	6.48	2	< 0.05	4.04	2	0.13

	(e) /ɑ/			(f) /ʌ/		
	χ2	*df*	*p*-value	*χ2*	*df*	*p*-value
(Intercept)	0.03	1	0.86	17.39	1	< 0.001
Attempt	9.06	2	< 0.05	13.18	2	< 0.01
Condition	1.20	1	0.27	7.40	1	< 0.01
Attempt x Condition	1.94	2	0.38	4.16	2	0.13

In the following section, we examine whether speakers modified their speech across attempts and conditions by analyzing the acoustic characteristics of their productions, including vowel duration, F1, and F2.

3.2 Analysis 2: Acoustic features

3.2.1 Duration

We first examined how vowel duration varied as a function of ATTEMPT and CONDITION. Figure 2 shows vowel duration across attempts and conditions, and Table 2 presents the mixed-effects regression results for duration for each of the six vowel categories. In Figures 2–4, the “First” category includes only initially

misrecognized tokens and serves as the baseline, whereas the "Second" and "Third" categories include all subsequent productions of those same initially misrecognized words, regardless of whether they were recognized correctly. This approach was adopted to capture acoustic changes relative to the initial misrecognized production, allowing us to track how learners adjusted their pronunciation across repeated attempts following an initial misrecognition. For comparison, these figures also present "Correct," which denotes tokens that were correctly recognized on the first attempt.

The effect of ATTEMPT on vowel duration was significant for four vowel categories: /i/ (Table 2(a)), /ɛ/ (Table 2(c)), /ɑ/ (Table 2(e)), and /ʌ/ (Table 2(f)), with no reliable effects for the remaining vowels. Pairwise comparisons using the *emmeans* function were conducted to identify which attempts differed significantly in duration. For all four vowels, the third attempt was significantly longer than the first (all $p < 0.05$). /i/ also exhibited a significant increase from the first to the second attempt ($p < 0.001$). No significant differences were observed between the second and third attempts for any of the vowels. The ATTEMPT × CONDITION interaction was significant for /ɑ/, and follow-up pairwise comparisons indicated that the lengthening from the first to third attempt occurred in the Google Docs condition only ($p < 0.001$). The effect of CONDITION was significant for four vowels /i, ɪ, ɑ, ʌ/ and marginally significant for /ɛ, æ/, with consistently longer duration in the Google Docs condition than in the Alexa condition.

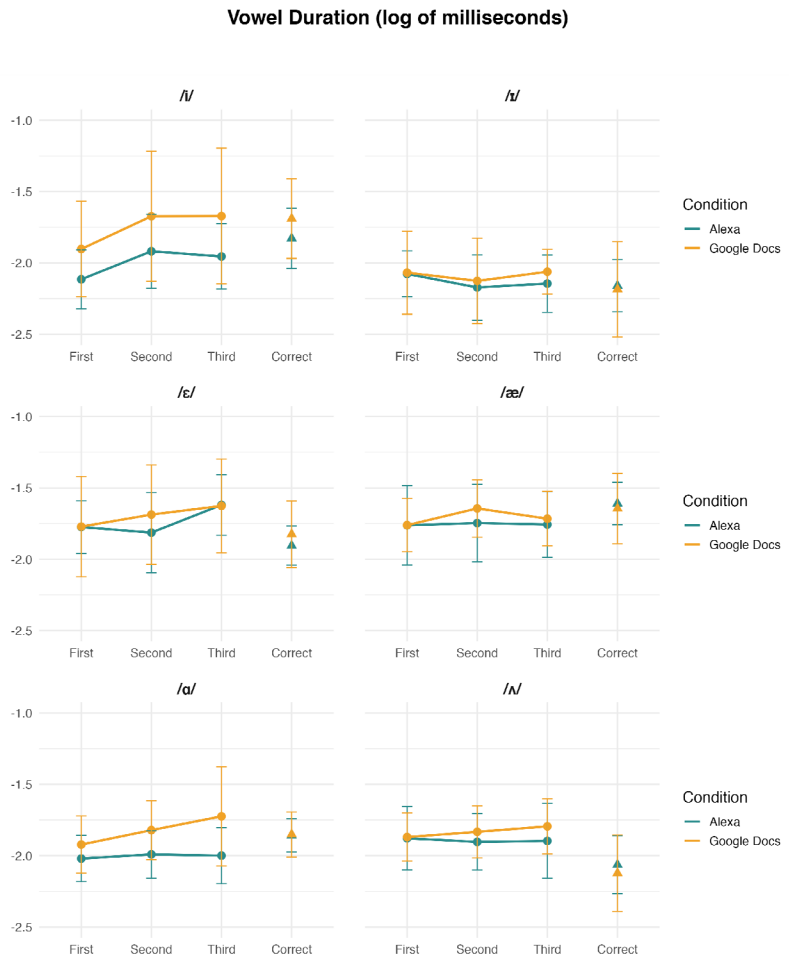

Figure 2. Vowel durations of the six vowels across attempts, separated by condition. Error bars represent standard deviations.

Table 2. Results of statistical analyses examining the effects of ATTEMPT and CONDITION on vowel duration

	(a) /i/			(b) /ɪ/		
	χ2	*df*	*p*-value	*χ2*	*df*	*p*-value
(Intercept)	486.15	1	< 0.001	1397.91	1	< 0.001
Attempt	19.41	2	< 0.001	3.49	2	0.17
Condition	14.30	1	< 0.001	6.87	1	< 0.01
Attempt x Condition	0.77	2	0.68	1.52	2	0.47

	(c) /ɛ/			(d) /æ/		
	χ2	*df*	*p*-value	*χ2*	*df*	*p*-value
(Intercept)	462.59	1	< 0.001	674.51	1	< 0.001
Attempt	15.60	2	< 0.001	5.06	2	0.08
Condition	3.48	1	0.06	3.04	1	0.08
Attempt x Condition	3.06	2	0.22	2.25	2	0.32

	(e) /ɑ/			(f) /ʌ/		
	χ2	*df*	*p*-value	*χ2*	*df*	*p*-value
(Intercept)	758.46	1	< 0.001	1143.68	1	< 0.001
Attempt	9.89	2	< 0.01	7.25	2	< 0.05
Condition	31.02	1	< 0.001	13.87	1	< 0.001
Attempt x Condition	8.31	2	< 0.05	2.61	2	0.27

3.2.2 F1

Figure 3 shows F1 across ATTEMPT and CONDITION for the six vowels, and Table 3 presents the mixed-effects regression results for F1. As in the duration analyses above, each subpanel (a-f) shows the results for the six vowels in the same order: /i/, /ɪ/, /ɛ/, /æ/, /ɑ/, and /ʌ/. As shown in Table 3, no significant effects of ATTEMPT, CONDITION, or their interaction were found for F1 across vowels.

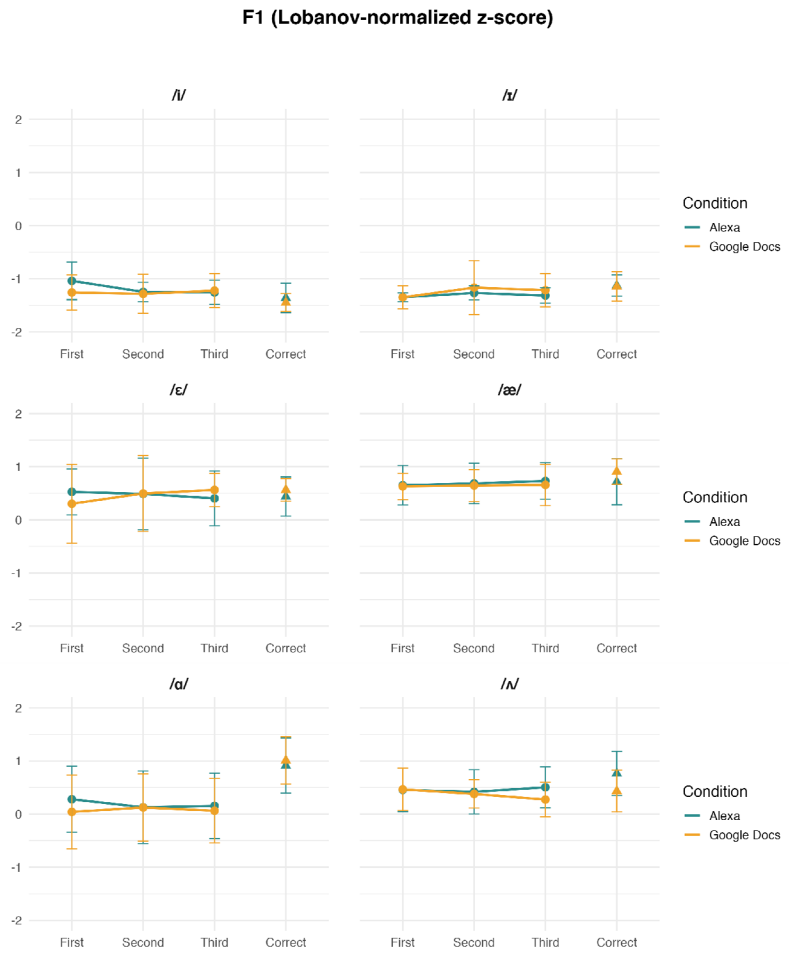

Figure 3. F1 values of the six vowels across attempts, separated by condition. Error bars represent standard deviations.

Table 3. Results of statistical analyses examining the effects of ATTEMPT and CONDITION on F1

	(a) /i/			(b) /ɪ/		
	$\chi2$	*df*	*p*-value	$\chi2$	*df*	*p*-value
(Intercept)	492.50	1	< 0.001	384.36	1	< 0.001
Attempt	1.00	2	0.61	4.12	2	0.13
Condition	0.06	1	0.81	1.47	1	0.23
Attempt x Condition	1.79	2	0.41	1.28	2	0.53

	(c) /ɛ/			(d) /æ/		
	$\chi2$	*df*	*p*-value	$\chi2$	*df*	*p*-value
(Intercept)	23.79	1	< 0.001	116.80	1	< 0.001
Attempt	0.20	2	0.91	0.98	2	0.61
Condition	1.05	1	0.31	0.15	1	0.70
Attempt x Condition	0.65	2	0.72	0.02	2	0.99

	(e) /ɑ/			(f) /ʌ/		
	$\chi2$	*df*	*p*-value	$\chi2$	*df*	*p*-value
(Intercept)	9.01	1	< 0.01	31.00	1	< 0.001
Attempt	3.89	2	0.14	3.84	2	0.15
Condition	1.10	1	0.29	3.31	1	0.07
Attempt x Condition	0.55	2	0.76	2.41	2	0.30

3.2.3 F2

The effect of ATTEMPT on F2 was observed only for /ɛ/ (Table 4(c)) and /ʌ/ (Table 4(f)). Pairwise comparisons indicated a decrease in F2 values from the second to the third attempt ($p < 0.05$) for /ɛ/ and from the first to the third attempt ($p < 0.01$) for /ʌ/. No reliable differences were found between the other attempt pairs. The effect of CONDITION on F2 was significant for /ɪ/, /ɛ/, and /ʌ/ (see Tables 4(b), 4(c), and 4(f), respectively). For /ɪ/ and /ʌ/, F2 values were higher in the Alexa condition than in the Google Docs condition (see Figure 4). For /ɛ/, the opposite pattern was found, with higher F2 values in the Google Docs condition than in the Alexa condition (see Figure 4). However, as indicated by the significant ATTEMPT × CONDITION

interaction, this difference was primarily observed at the third attempt ($p < 0.01$).

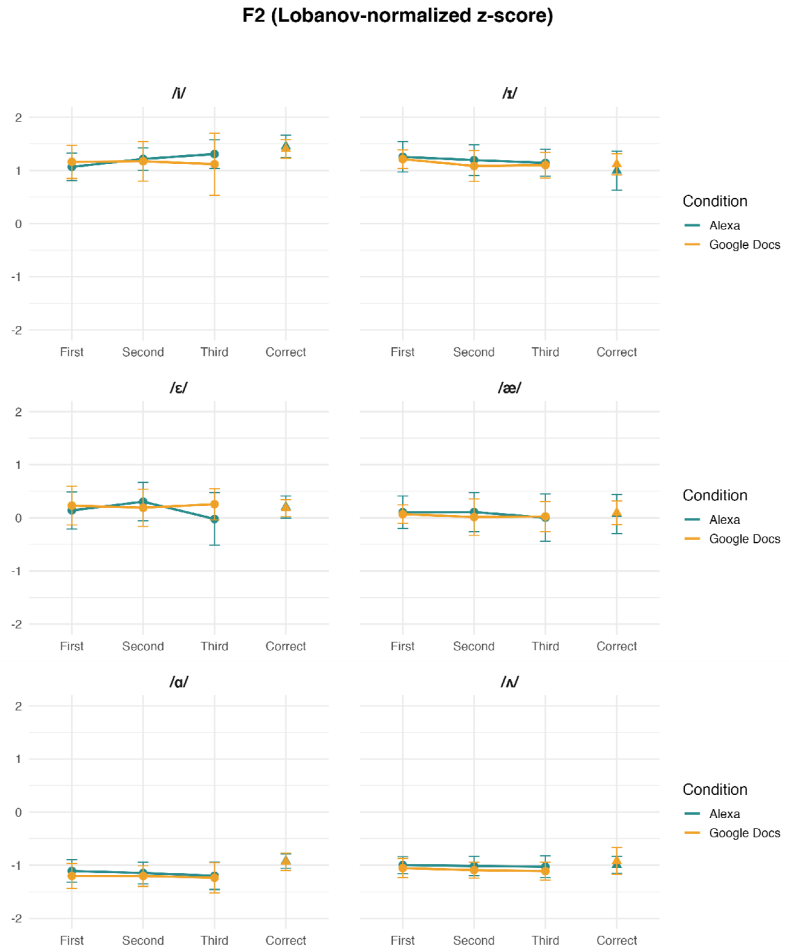

Figure 4. F2 values of the six vowels across attempts, separated by condition. Error bars represent standard deviations.

Table 4. Results of statistical analyses examining the effects of ATTEMPT and CONDITION on F2

	(a) /i/			(b) /ɪ/		
	χ2	*df*	*p*-value	*χ2*	*df*	*p*-value
(Intercept)	697.96	1	< 0.001	392.44	1	< 0.001
Attempt	3.81	2	0.15	1.65	2	0.44
Condition	0.95	1	0.33	4.18	1	< 0.05
Attempt x Condition	0.15	2	0.93	2.72	2	0.26

	(c) /ɛ/			(d) /æ/		
	χ2	*df*	*p*-value	*χ2*	*df*	*p*-value
(Intercept)	8.69	1	< 0.01	0.37	1	0.54
Attempt	7.36	2	< 0.05	3.59	2	0.17
Condition	5.35	1	< 0.05	0.43	1	0.51
Attempt x Condition	13.24	2	< 0.01	1.28	2	0.53

	(e) /ɑ/			(f) /ʌ/		
	χ2	*df*	*p*-value	*χ2*	*df*	*p*-value
(Intercept)	141.32	1	< 0.001	373.11	1	< 0.001
Attempt	0.29	2	0.87	11.01	2	< 0.01
Condition	2.07	1	0.15	6.57	1	< 0.05
Attempt x Condition	0.04	2	0.98	0.52	2	0.77

To summarize the acoustic results, vowel duration was most consistently affected by ATTEMPT and CONDITION. Vowel duration was generally longer in the third attempt than in the first, and longer in the Google Docs condition than in the Alexa condition. For F1, no significant effects of ATTEMPT, CONDITION, or their interaction were observed, and all vowels remained stable across attempts and conditions. For F2, ATTEMPT effects emerged for /ɛ/ and /ʌ/, with decreased F2 values in the third attempt. CONDITION effects were observed for /ɪ/, /ɛ/, and /ʌ/; /ɪ/ and /ʌ/ showed higher F2 in the Alexa condition, whereas /ɛ/ showed higher F2 in the Google Docs condition, with this difference emerging primarily at the third attempt.

4. Discussion

This study examined how Korean learners of English adjusted their vowel production when interacting with two types of ASR systems: an AI assistant (Amazon Alexa) and a STT tool (Google Docs). The study addressed three primary questions: First, does the word recognition accuracy of L2 learners improve over the course of multiple pronunciation attempts following feedback from the two ASR systems? Second, when recognition fails, do learners modify the acoustic properties of their vowels (duration, F1, and F2) across subsequent attempts relative to their initial misrecognized production? Third, how do differences between the two ASR systems influence L2 learners' pronunciation adjustments?

The current study yielded three main findings, each addressing the research questions outlined above. First, L2 learners' word recognition accuracy improved across successive pronunciation attempts. This pattern suggests that learners were able to notice and respond to the feedback from both ASR systems, although the improvement may also partly reflect other factors, such as ASR system adaptation. Second, when a misrecognition occurred, participants modified their subsequent productions. These modifications were most consistently reflected in increased vowel duration, while changes in spectral properties were limited. There are several potential explanations for the observed increase in vowel duration, which we return to later in the discussion. Third, the two ASR systems appeared to elicit somewhat different patterns of acoustic adjustment, which may reflect differences in how learners approached each system during the task. Learners tended to treat Alexa more like an interactive conversational partner and Google Docs more like a transcription tool requiring careful enunciation, although these interpretations should be considered tentative given the limited statistical support. These findings are discussed in more detail below.

For most vowels examined, recognition accuracy showed a robust ATTEMPT effect; the second and third attempts were recognized more accurately than the first attempts, while the accuracy of the second and third attempts did not differ significantly. In other words, learners made their largest improvement immediately after the first misrecognition, after which their performance stabilized. This sharp increase between the first and second attempts, compared to the plateauing trend thereafter, suggests that the initial repair is the most effective window for adjustment.

The finding also indicates that L2 learners were systematically modifying their productions in response to the ASR systems' feedback rather than simply repeating the initial token. Overall, the findings are consistent with prior research on technology-assisted pronunciation learning, which shows that repeated, feedback-rich practice with ASR can lead to reliable gains in segmental accuracy (Bashori et al. 2024; Ngo et al. 2024; Amrate and Tsai 2025).

Despite the significant improvement in recognition accuracy observed in later attempts, the acoustic analyses showed that the adjustments learners made across attempts were modest. In particular, changes in F1 and F2 were small and limited to only a few vowels, indicating that learners were not systematically shifting their productions in the F1-F2 vowel space. Instead, participants primarily lengthened the vowel when attempting to correct their pronunciation. This reliance on duration is consistent with findings from cross-language speech perception research showing that L2 learners often rely on duration cues, unlike native English listeners who rely predominately on spectral cues, with vowel duration serving as a secondary cue (Escudero and Boersma 2002; Morrison 2002; Kim et al. 2018). Similarly, Bohn (1995) argued that when L2 learners' prior linguistic experience does not sensitize them to spectral differences, they tend to rely on duration cues as the primary means of differentiating difficult vowel contrast. In the present study, short-term adjustments following ASR feedback did not involve systematic shifts in spectral vowel targets, extending previous L2 research by suggesting that learners' reliance on temporal cues may persist even in ASR-mediated pronunciation practice.

The difficulty of making precise spectral adjustments may have been increased by the type of feedback provided by the two ASR systems. Although both Alexa and Google Docs reliably indicated when a learner's pronunciation was misrecognized, neither system supplied information about how the pronunciation should be corrected. Learners could hear or see how their production had been recognized, often as a different English word, but they were not given any guidance about the specific articulatory or acoustic changes required to produce the target vowel accurately. For low-proficiency learners whose phonetic knowledge of English vowel categories is limited, the absence of corrective feedback may pose a substantial challenge. Without access to a native-like model or explicit feedback about which dimension (height, backness, tenseness, rounding) needed adjustment, learners may have been unsure how to move their vowel in the F1–F2 space.

Regarding the effects of CONDITION (i.e., Alexa vs. Google Docs), some ATTEMPT × CONDITION interactions reached significance, although this pattern was not consistent across all vowels and measures, suggesting that adjustment differences between the two systems emerged only in specific contexts. It is important to note that the conditions also differed procedurally. In the Alexa condition, participants produced the phrase "Echo, spell ____," embedding the target word within a short carrier sentence, whereas in the Google Docs condition they activated the STT tool, produced a single isolated word, and then deactivated recognition, thus creating discrete, isolated-word dictation tasks. These procedure differences likely influenced production patterns: Alexa encouraged more phrase-level, conversational production, while Google Docs promoted careful, item-focused articulation. The acoustic consequences of these differences were most evident in vowel duration, with vowels consistently longer in the Google Docs condition. Importantly, this design means that the observed CONDITION effects may reflect not only differences between the ASR systems, but also differences in speaking task (phrase-embedded vs. isolated word production). Therefore, the results should be interpreted with caution, and the CONDITION effect should not be attributed solely to differences between the two systems.

These procedural contrasts may also have shaped participants' perceptions of the systems. Google Docs' button-activated transcription and visual feedback can position it as a "tool" or "scoring device," whereas Alexa, addressed as "Echo" with spoken replies, may invite a more conversational stance. These interpretations are in line with prior research on ASR-equipped learning websites and voice assistants, where learners often perceived such systems either as evaluators or as conversational partners and adapted their speech accordingly (Neri 2007; Purington et al. 2017; Cohn and Zellou 2021; Fortunati et al. 2022; Bashori et al. 2024).

One notable observation from the present data is the lack of a clear one-to-one correspondence between ASR recognition accuracy and the specific acoustic adjustments made by learners. For example, vowel duration was generally longer in the Google Docs condition, even when recognition accuracy was similar or lower than in the Alexa condition. This mismatch suggests that learners were likely adjusting multiple acoustic cues simultaneously rather than relying on a single dimension. That is, small, distributed changes across several acoustic dimensions may have combined to produce speech signals that were closer to the target category, cumulatively crossing

the ASR decision boundary. This interpretation is consistent with phonetic evidence showing that perceptual judgments often reflect holistic integration of multiple parameters rather than straightforward mappings from individual acoustic measures. For example, judgments of phonetic alignment or convergence are frequently holistic; listeners integrate multiple cues to form a global percept, and perceptual ratings do not necessarily correlate in a simple way with any single acoustic dimension (Pardo 2006; Babel and Bulatov 2012; Pardo 2013). From this perspective, ASR systems may likewise rely on multiple acoustic cues in combination, such that recognition outcomes reflect the cumulative effect of incremental adjustments across dimensions. The present study focused on individual acoustic cues, providing a useful first step toward understanding how learners modify their speech in response to ASR feedback. Future work could build on this foundation by examining how multiple acoustic cues jointly contribute to ASR recognition.

These findings have practical implications for designing more effective ASR-based pronunciation activities for classroom L2 learners. First, our findings show that learners predominantly relied on vowel duration, even when ASR recognition scores improved. Instructors should therefore interpret high recognition rates with caution and, when needed, supplement them with explicit assessment and training focused on vowel quality. Second, voice-activated AI interfaces may offer opportunities to frame pronunciation practice as interactive rather than test-like. Such interactional formats could encourage learners to produce speech that is more natural and balanced across multiple acoustic dimensions, rather than relying solely on exaggerated vowel duration to achieve recognition. Third, our findings indicate that recognition accuracy improved substantially from the first attempt to the second, but then stabilized thereafter. This plateau may result from the absence of a clear model pronunciation for learners to follow. Providing a target pronunciation, whether from a teacher or an ASR system, could support further improvement by giving learners a concrete reference for adjusting both temporal and spectral aspects of their speech. Finally, the results support recent recommendations in Computer-Assisted Pronunciation Training (CAPT) research to move beyond isolated segment scores and incorporate listener-based outcomes, such as intelligibility and comprehensibility, to capture meaningful improvements in pronunciation (Amrate and Tsai 2025).

While the present study provides insights into ASR-based pronunciation practice, several limitations suggest directions for future research. First, one key caveat concerns

the underlying recognition mechanisms of the two ASR systems. Because both Amazon Alexa and Google Docs STT are proprietary systems, detailed information about how they weight specific acoustic cues (e.g., formant frequencies, duration) is not publicly available. Contemporary ASR systems typically integrate acoustic and language modeling components, allowing recognition to be shaped not only by phonetic detail but also by lexical frequency, contextual constraints, interaction history, and platform-specific limitations on permissible utterances (Amazon Science 2021; Google Cloud Speech-to-Text no date). Accordingly, the relative contribution of acoustic versus contextual information may differ across platforms and interaction formats.

In the present study, efforts were made to minimize cumulative contextual prediction and maintain consistency across trials. In the Alexa condition, target words were embedded within a fixed carrier phrase, whereas in the Google Docs condition, they were produced in isolation and the system was reset across trials to maintain comparable transcription conditions. Nevertheless, the internal contribution of acoustic versus linguistic modeling cannot be directly estimated within this design.

The observed differences in learners' acoustic adjustments across systems may therefore reflect not only differences in how learners approached the task but also differences in how each ASR system integrates acoustic and contextual information. However, this limitation makes it difficult to determine the exact source of improvement across repeated attempts. Although recognition accuracy increased over trials, such improvement cannot be attributed solely to learners' intentional phonetic adjustment. Alternative explanations include short-term speaker adaptation or decoding stabilization within the ASR system (Lin et al. 2024) or a combination of learner- and system-level factors.

Importantly, recent research has shown that ASR feedback can enhance learners' awareness of pronunciation errors and lead to improvements in pronunciation as evaluated by human listeners (Dai and Wu 2023). In light of these findings, the observed increase in ASR recognition accuracy in the present study is at least consistent with the possibility of pronunciation improvement, although the current design does not allow for a definitive causal interpretation. Future research could more systematically manipulate acoustic cues across trials and track recognition outcomes in order to isolate learner-driven adjustments from ASR-internal processes.

Second, because the study did not include a native-speaker control condition, it remains unclear whether ASR feedback was sufficiently accurate to function as a

diagnostic tool comparable to native listeners or teachers. Nevertheless, prior work suggests that the systems used here perform broadly in line with human judgments, at least for English vowels and related tasks. For Alexa, Song et al. (2022) reported recognition rates comparable to native-speaker performance for several English vowel categories. For Google Docs' STT, Hirai and Kovalyova (2023) showed that STT outcomes aligned with native-like evaluators' pronunciation ratings, with strong interrater reliability supporting this correspondence. At the same time, other work has reported that mobile ASR dictation output was comparable to native listeners for some vowels (e.g., /i, æ, u/), but differed by approximately 10–20 percentage points for others (e.g., /ɪ, ɛ, ɑ, ʌ/) (Guskaroska 2020). Taken together, these findings support that ASR feedback can be regarded as a reasonably informative diagnostic tool, while also recognizing that its reliability may vary across vowel categories and systems.

Third, in the present study, L2 proficiency was assessed solely through a self-reported CEFR level, which may be susceptible to subjective bias and may not fully capture domain-specific language skills (e.g., listening, speaking, writing). Future research should incorporate objective proficiency measures (e.g., TOEFL/TOEIC scores) to characterize participants' L2 level more rigorously and to examine how domain-specific language ability relates to phonetic adjustment patterns. Furthermore, future research should recruit learners across a broader range of proficiency to examine whether L2 level shapes how systematically speakers adjust spectral cues (F1/F2) after ASR feedback. The limited proficiency range in the present sample makes it difficult to evaluate proficiency-related differences in adjustment patterns. It may understate the extent to which more advanced learners fine-tune formant targets beyond mainly temporal changes.

Fourth, although vowel duration increased across attempts for several vowels, the present analyses cannot determine whether learners specifically lengthened the target vowel or instead slowed their overall speech rate following misrecognition. Participants may have adopted a more careful speaking style in later attempts, resulting in longer vowel durations without vowel-specific adjustment. Because overall speech rate was not measured, this possibility cannot be ruled out and should be taken into consideration in interpreting the results.

Fifth, the current acoustic analyses focused on duration, F1, and F2, even though both human listeners and ASR systems are known to draw on a broader set of segmental and suprasegmental cues such as spectral tilt, voice quality, stress patterns,

and global prosody (Munro and Derwing 1995; Harrington 2010). Extending analyses to these dimensions could provide a more complete picture of how learners adjust their speech after misrecognition, and what cues ASR systems use in combination when mapping learner productions onto target categories. Addressing these limitations will enable future research to build upon the present findings and further inform the design of effective ASR-based pronunciation practice.

References

Amazon Science. 2021. The engineering behind Alexa's contextual speech recognition. Retrieved Feb. 15, 2026 from https://www.amazon.science/latest-news/the-engineering-behind-alexas-contextual-speech-recognition.

Amrate, Moustafa and Pi-hua Tsai. 2025. Computer-assisted pronunciation training: A systematic review. *ReCALL* 37(1): 22–42.

Babel, Molly and Dasha Bulatov. 2012. The role of fundamental frequency in phonetic accommodation. *Language and Speech* 55(2): 231–248.

Baran-Łucarz, Małgorzata and Jang Ho Lee. 2021. Selected determinants of pronunciation anxiety. *International Journal of English Studies* 21(1): 93–113.

Barr, Dale J., Roger Levy, Christoph Scheepers, and Harry J. Tily. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language* 68(3): 255–278.

Bashori, Muzakki, Roeland van Hout, Helmer Strik, and Catia Cucchiarini. 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.

Bates, Douglas, Martin Mächler, Benjamin Bolker, and Steven Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67: 1–48.

Bohn, Ocke-Schwen and James E. Flege. 1995. The perception of foreign and native vowels by L2 learners. In Allan James and Jonathan Leather (eds.), *The acquisition of phonology and phonetics,* 115–125. Cambridge: Cambridge University Press.

Cohn, Michelle and Georgia Zellou. 2021. Prosodic differences in human-and Alexa-directed speech, but similar local intelligibility adjustments. *Frontiers in Communication* 6: 675704.

Couper, Graeme. 2021. Pronunciation teaching issues: Answering teachers' questions. *RELC Journal* 52(1): 128–143.

Cucchiarini, Catia and Helmer Strik. 2017. Automatic speech recognition for second language pronunciation training. In Okim Kang and Ron I. Thomson (eds.), *The Routledge handbook of contemporary English pronunciation,* 556–569. London: Routledge.

Cucchiarini, Catia, Warda Nejjari, and Helmer Strik. 2012. My pronunciation coach: Improving English pronunciation with an automatic coach that listens. *Language Learning in Higher Education* 1(2): 365–376.

Dai, Yuanjun and Zhiwei Wu. 2023. Mobile-assisted pronunciation learning with feedback from peers and/or automatic speech recognition: A mixed-methods study. *Computer Assisted Language Learning* 36(5–6): 861–884.

Dizon, Gilbert. 2017. Using intelligent personal assistants for second language learning: A case study of Alexa. *TESOL Journal* 8(4): 811–830.

Dizon, Gilbert. 2020. Evaluating intelligent personal assistants for L2 listening and speaking development. *Language Learning & Technology* 24(1): 16–26.

Dizon, Gilbert. 2023. Affordances and constraints of intelligent personal assistants for second-language learning. *RELC Journal* 54(3): 848–855.

Escudero, Paola and Paul Boersma. 2002. The subset problem in L2 perceptual development: Multiple-category assimilation by Dutch learners of Spanish. In Barbora Skarabela, Sarah Fish, and Anna Do (eds.), *Proceedings of the 26th Annual Boston University Conference on Language Development,* 208–219. Somerville, MA: Cascadilla Press.

Evers, Katerina and Sufen Chen. 2022. Effects of an automatic speech recognition system with peer feedback on pronunciation instruction for adults. *Computer Assisted Language Learning* 35(8): 1869–1889.

Flege, James E., Ocke-Schwen Bohn, and Sunyoung Jang. 1997. Effects of experience on non-native speakers' production and perception of English vowels. *Journal of Phonetics* 25(4): 437–470.

Fortunati, Leopoldina, Autumn Edwards, Chad Edwards, Anna Maria Manganelli, and Federico De Luca. 2022. Is Alexa female, male, or neutral? A cross-national and cross-gender comparison of perceptions of Alexa's gender and status as a communicator. *Computers in Human Behavior* 137: 107426.

Fox, John and Sanford Weisberg. 2018. *An R companion to applied regression.* Thousand Oaks, CA: Sage.

Google Cloud Speech-to-Text. no date. Recognition Config, Google Cloud documentation. Retrieved Feb. 15, 2026 from https://docs.cloud.google.com/speech-to-text/docs/reference/rest/v1/RecognitionConfig

Guskaroska, Agata. 2020. ASR-dictation on smartphones for vowel pronunciation practice. *Journal of Contemporary Philology* 3(2): 45-61.

Gutz, Sarah E., Marc F. Maffei, and Jordan R. Green. 2023. Feedback from automatic speech recognition to elicit clear speech in healthy speakers. *American Journal of Speech-language Pathology* 32(6): 2940–2959.

Harrington, Jonathan. 2010. *Phonetic analysis of speech corpora.* Oxford: Wiley-Blackwell.

Hirai, Akiyo and Angelina Kovalyova. 2023. Using speech-to-text applications for assessing English language learners' pronunciation: A comparison with human raters. In

Maria-del-Mar Suárez and Walaa M. El-Henawy (eds.), *Optimizing online English language learning and teaching,* 337–355. Cham: Springer International Publishing.

Inceoglu, Solene. 2022. Developing pronunciation learner autonomy with automatic speech recognition and shadowing. In Shannon McCrocklin (ed.), *Technological resources for second language pronunciation learning and teaching: Research-based approaches,* 171–192. London: Lexington Books.

Kim, Donghyun, Meghan Clayards, and Heather Goad. 2018. A longitudinal study of individual differences in the acquisition of new vowel contrasts. *Journal of Phonetics* 67: 1–20.

Leis, Adrian. 2025. How speech-to-text technology affects pronunciation gains and self-confidence in EFL learners. *Computer Assisted Language Learning* 1–24.

Lenth, Russell. 2022. *emmeans: Estimated marginal means, aka least-squares means.* R package version 1.7.2.

Levis, John M. 2022. Teaching pronunciation: Truths and lies. In Camilla Bardel, Christina Hedman, Katarina Rejman, and Elisabeth Zetterholm (eds.), *Exploring language education,* 39-72. Stockholm: Stockholm University Press.

Liakin, Denis, Walcir Cardoso, and Natalia Liakina. 2017. Mobilizing instruction in a second-language context: Learners' perceptions of two speech technologies. *Languages* 2(3): 11.

Lin, Guan-Ting, Wei Ping Huang, and Hung-yi Lee. 2024. Continual test-time adaptation for end-to-end speech recognition on noisy speech. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing,* 20003–20015. Maiami, FL: Association for Computational Linguistics.

Matuschek, Hannes, Reinhold Kliegl, Shravan Vasishth, Harald Baayen, and Douglas Bates. 2017. Balancing Type I error and power in linear mixed models. *Journal of Memory and Language* 94: 305–315.

McCrocklin, Shannon M. 2016. Pronunciation learner autonomy: The potential of automatic speech recognition. *System* 57(1): 25–42.

Miller, Krista A., Gary E. Raney, and Alexander P. Demos. 2020. Time to throw in the towel? No evidence for automatic conceptual metaphor access in idiom processing. *Journal of Psycholinguistic Research* 49(5): 885–913.

Morrison, Geoffrey S. 2002. Japanese listeners' use of duration cues in the identification of English high front vowels. In Julie Larson and Mary Paster (eds.), *Proceedings of the 28th Annual Meeting of the Berkeley Linguistics Society,* 189–200. Berkeley, CA: Berkeley Linguistic Society.

Moussalli, Souheila and Walcir Cardoso. 2021. Intelligent personal assistants and L2 pronunciation development: Focus on English past-*ed*. In Naouel Zoghlami, Cédric Brudermann, Cédric Sarré, Muriel Grosbois, Linda Bradley, and Sylvie Thouësny (eds.), *CALL and Professionalisation: Short Papers from EUROCALL 2021*, 226-231. Paris, France: Research-publishing.net.

Munro, Murray J. and Tracey M. Derwing. 1995. Processing time, accent, and comprehensibility in the perception of native and foreign-accented speech. *Language and Speech* 38(3): 289–306.

Neri, Ambra. 2007. *The pedagogical effectiveness of ASR-based computer assisted pronunciation training*. PhD Dissertation. Radboud University.

Ngo, Thuy Thi-Nhu, Howard Hao-Jan Chen, and Kyle Kuo-Wei Lai. 2024. The effectiveness of automatic speech recognition in ESL/EFL pronunciation: A meta-analysis. *ReCALL* 36(1): 4–21.

Pardo, Jennifer S. 2006. On phonetic convergence during conversational interaction. *Journal of the Acoustical Society of America* 119(4): 2382–2393.

Pardo, Jennifer S. 2013. Measuring phonetic convergence in speech production. *Frontiers in Psychology* 4: 559.

Purington, Amanda, Jessie G. Taft, Shruti Sannon, Natalya N. Bazarova, and Samuel H. Taylor. 2017. "Alexa is my new BFF": Social roles, user satisfaction, and personification of the Amazon Echo. In Gloria Mark and Susan Fussell (eds.), *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems,* 2853–2859. New York, NY: Association for Computing Machinery.

Shortt, Mitchell, Shantanu Tilak, Irina Kuznetcova, Bethany Martens, and Babatunde Akinkuolie. 2023. Gamification in mobile-assisted language learning: A systematic review of Duolingo literature from public release of 2012 to early 2020. *Computer Assisted Language Learning* 36(3): 517–554.

Song, Jae Yung, Anne Pycha, and Tessa Culleton. 2022. Interactions between voice-activated AI assistants and human speakers and their implications for second-language acquisition. *Frontiers in Communication* 7: 995475.

Tsang, Art. 2025. The relationships between EFL learners' anxiety in oral presentations, self-perceived pronunciation, and speaking proficiency. *Language Teaching Research* 29(4): 1639–1659.

Yang, Byunggon. 1996. A comparative study of American English and Korean vowels produced by male and female speakers. *Journal of Phonetics* 24(2): 245–261.

Yang, Christine Ting-Yu, Shu-Li Lai, and Howard Hao-Jan Chen. 2024. The impact of intelligent personal assistants on learners' autonomous learning of second language listening and speaking. *Interactive Learning Environments* 32(5): 2175–2195.

Appendix

A. Target words

/i/	beat, peak, seek, deep, seat, heat
/ɪ/	bit, pick, sick, dip, sit, hit
/ɛ/	bet, met, pet, set, guess, mess
/æ/	bat, mat, pat, sat, gas, mass
/ɑ/	cop, not, dock, hot, shot, got
/ʌ/	cup, nut, duck, hut, shut, gut

B. Filler words

/eɪ/	fake, bake, mate, great
/aɪ/	fight, bike, kite, might
/oʊ/	goat, boat, coat, note

Liying Bai
Graduate Student
Department of English Language and Literature
Chung-Ang University
84, Heukseok-ro, Dongjak-gu,
Seoul, 06974, Republic of Korea
E-mail: baek0904@cau.ac.kr

Jae Yung Song
Professor
Department of English Language and Literature
Chung-Ang University
84, Heukseok-ro, Dongjak-gu,
Seoul, 06974, Republic of Korea
E-mail: songjy@cau.ac.kr

Received: 2026. 01. 19.
Revised: 2026. 02. 26.
Accepted: 2026. 02. 27.