frontiers | Frontiers in Communication

TYPE Original Research
PUBLISHED 21 October 2022
DOI 10.3389/fcomm.2022.995475

OPEN ACCESS

EDITED BY
Marcel Pikhart,
University of Hradec Králové, Czechia

REVIEWED BY
Alejandro Javier Wainselboim,
CONICET Mendoza, Argentina
Gilbert Dizon,
Himeji Dokkyo University, Japan
Olga Dmitrieva,
Purdue University, United States

*CORRESPONDENCE
Jae Yung Song
songjy@cau.ac.kr

SPECIALTY SECTION
This article was submitted to Language Sciences, a section of the journal Frontiers in Communication

RECEIVED 15 July 2022
ACCEPTED 05 October 2022
PUBLISHED 21 October 2022

CITATION
Song JY, Pycha A and Culleton T (2022) Interactions between voice-activated AI assistants and human speakers and their implications for second-language acquisition. *Front. Commun.* 7:995475. doi: 10.3389/fcomm.2022.995475

Interactions between voice-activated AI assistants and human speakers and their implications for second-language acquisition

Jae Yung Song[1,2]*, Anne Pycha[2] and Tessa Culleton[2]

[1]Department of English Language and Literature, Chung-Ang University, Seoul, South Korea, [2]Department of Linguistics, University of Wisconsin-Milwaukee, Milwaukee, WI, United States

Voice-activated artificially intelligent (voice-AI) assistants, such as Alexa, are remarkably effective at processing spoken commands by native speakers. What happens when the command is produced by an L2 speaker? In the current study, we focused on Korean-speaking L2 learners of English, and we asked (a) whether Alexa could recognize intended productions of two vowel contrasts, /i/ vs. /ɪ/ and /æ/ vs. /ɛ/, that occur in English but not in Korean, and (b) whether L2 talkers would make clear-speech adjustments when Alexa misrecognized their intended productions. L2 talkers ($n = 10$) and native English ($n = 10$) controls asked Alexa to spell out words. Targets were words that formed minimal vowel pairs, e.g., *beat-bit, pet-pat*. Results showed that Alexa achieved a 55% accuracy rate with L2 productions, compared to 98% for native productions. When Alexa misrecognized an intended production (e.g., spelling *P-E-T* when the speaker intended *pat*), L2 talkers adjusted their subsequent production attempts by altering the duration, F1 and F2 of individual vowels (except for /ɛ/), as well as increasing vowel duration difference between contrasting vowels. These results have implications for theories of speech adaptation, and specifically for our understanding of L2 speech modifications oriented to voice-AI devices.

KEYWORDS

human-computer interaction, clear speech, L2 acquisition, vowel recognition, voice-AI

Introduction

Voice-activated artificially intelligent (voice-AI) assistants, such as Google Assistant, Apple's Siri, and Amazon's Alexa, respond to spoken human questions using natural-sounding speech generated by computer algorithms. With the widespread use of voice-AI, there are ever-increasing ways we use our voices to interact with our world, such as converting speech to text or asking questions without typing. However, we are only beginning to understand interactions between voice-AI and human speakers, and second-language (L2) learners in particular. As the first step

to better understand the challenges that L2 learners might face in using voice-AI in the target language, it is important to examine how their verbal interaction with voice-AI compares with that of native speakers. Thus, in this study, we examined L2 learners' productions when giving commands to a voice-AI device, using words that contain vowel sounds that are non-existent in their phonemic inventory. In particular, we examined acoustic modifications L2 learners made when their initial word production was not correctly understood by voice-AI.

As a talker we constantly monitor the perceptual needs of the listener, and make various speech adaptations to accommodate them. For example, in degraded listening conditions, such as a noisy environment, we slow our speech down and articulate individual sounds more carefully (the "Lombard effect;" see Brumm and Zollinger, 2011). Similarly, when we talk to someone with a hearing impairment, we also make adjustments to our speech (e.g., Lindblom, 1990). When we make such adjustments, we use a special speech register called clear speech. The modifications in clear speech, which a talker adopts to be better understood, appear to translate into enhanced speech perception in listeners. Clear speech has been shown to improve intelligibility in various populations including normal-hearing adult listeners (Bradlow et al., 1996; Krause and Braida, 2002; Ferguson, 2004; Hazan et al., 2012), hard of hearing adults (Picheny et al., 1985), non-native adult listeners (Bradlow and Bent, 2002), infants (Song et al., 2010), and school-aged children with and without learning disabilities (Bradlow et al., 2003).

Production studies have suggested that clear speech is characterized by global properties such as slower speaking rate, increased loudness, increased fundamental frequency range, and more carefully articulated vowels as indicated by an expanded acoustic vowel space (Picheny et al., 1986; Krause and Braida, 2004; Smiljanić and Bradlow, 2005, 2009; Ferguson and Kewley-Port, 2007; Ferguson and Quené, 2014). In addition to these global changes, individual phonemes, especially vowels, also undergo change. For example, the direction of change for second formants (F2) from conversational to clear speech has been shown to depend on the front/back feature of the vowel. In studies by Ferguson and Kewley-Port (2002, 2007), F2 was shown to significantly increase for front vowels, whereas F2 showed no change or slight decreases for back vowels in clear speech. First formant (F1) values generally increased in clear speech across different vowels. Ferguson and Kewley-Port (2002) attributed the raised F1 to increased vocal effort in clear speech.

Studies on the acoustic properties of clear speech have traditionally focused on speech directed to various human populations who can benefit from enhanced intelligibility, including hard of hearing (Picheny et al., 1986), foreigners (Uther et al., 2007), and infants (Kuhl et al., 1997). Following Lindblom (1990) and others, we can make a general prediction that talkers will switch to the clear speech register based on their assessment of listener needs for any interlocutor. Interestingly, this prediction has also been extended to voice-AI (Uther et al., 2007): the basic idea is that listeners treat such voice-AI devices as if they require enhanced speech input. Recent findings have corroborated this prediction (Burnham et al., 2010; Cohn and Zellou, 2021; Cohn et al., 2021, 2022). For example, Cohn and Zellou (2021) examined the adjustments that speakers made in response to a misrecognition by a human or by voice-AI (Amazon's Alexa). They asked speakers to read sentences containing target words, and played back either human or Alexa pre-recorded responses. There were two potential responses for each target word (e.g., "*The word is bat*"): a "correctly understood" response ("*I think I heard bat*") and a "misrecognition" response ("*I'm not sure I understood. I think I heard bought or bat*"). Their results revealed some prosodic differences between Alexa-directed-speech and human-directed-speech, such as a decreased speech rate, higher mean F0, and greater F0 variation for Alexa-directed-speech. However, for sentences produced in response to a "misrecognition," speakers adjusted their speech similarly in both the Alexa and human conditions, exhibiting greater intensity, slower speech rate, higher mean F0, and greater F0 variation; in addition, back vowels were produced even further back (as indicated by lower F2), although there was no difference in F1. Thus, these results suggested that although there are some prosodic differences between Alexa-directed-speech and human-directed-speech, talker adjustments in response to misrecognition were similar for the two interlocutors, and were consistent with the qualities of clear speech.

While interactions between voice-AI and L2 learners do not yet constitute a full-fledged area of research, there are a number of studies which have explored these issues. For example, Dizon (2017) examined how accurately Alexa understood English utterances produced by L1 Japanese speakers. The results showed that, on average, Alexa understood only 50% of the learner commands. In Moussalli and Cardoso (2020), L2 speakers of various native languages were asked to interact with Alexa using a set of prepared questions. Although Alexa's accuracy rate in response to their utterances was relatively high (83%), it was still lower than the rate for human listeners (95%). In Dizon et al. (2022), Alexa understood ∼80% of the commands produced by L2 learners of Japanese. This study also reported that, in the face of a communication breakdown, listeners most commonly abandoned their attempt, rather than repeating their utterance or re-phrasing it. Other studies have also reported abandonment as a common response to breakdowns (Dizon and Tang, 2020; Tai and Chen, 2022) and noted that breakdowns are most commonly caused by mispronunciations, particularly for lower-level learners (Chen et al., 2020). Finally, Dizon (2020) reported that L2 English learners who interacted with Alexa made greater gains in L2 speaking proficiency than those who did not.

Most of these studies on the interactions between voice-AI and L2 learners have focused on the benefits of these

interactions for learning a foreign language. Through informal observations (Underwood, 2017) or a more structured survey or interview (Moussalli and Cardoso, 2016, 2020; Dizon, 2017), these studies suggested that voice-AI has the potential to support L2 acquisition by providing students opportunities to practice their L2 skills in a more enjoyable and engaging way. Furthermore, while Tai and Chen (2022) noted that communication breakdowns created anxiety in L2 learners because they doubted whether they had pronounced target words correctly, Dizon (2017) argued that by receiving indirect feedback, the L2 learners' attention was directed toward errors in their pronunciation, thereby encouraging them to correct these mistakes in subsequent exchanges. Although these studies have focused on how L2 learners' benefit from using voice-AI, to our knowledge, none have focused specifically on L2 learners' speech adaptations toward voice-AI.

With the prevalence and convenience of voice-AI, and based on research suggesting voice-AI's potential as a useful aid for language learning, it is only natural to seek a better understanding of the interaction between voice-AI and L2 learners. However, studies on the interaction between voice-AI and L2 learners are limited, and to our knowledge, there are no studies that provide systematic acoustic analyses of the speech adaptations of L2 learners. Yet we note that the L2 learning context provides a particularly authentic setting for studying voice-AI-oriented clear speech adaptations. In previous studies, in order to elicit clear speech directed to voice-AI, the authors presented native speakers with a misrecognition response even when their pronunciations were in fact appropriate (e.g., Cohn and Zellou, 2021). Unlike native speakers, L2 learners are more likely to experience a genuine communicative barrier when interacting with a voice-AI, especially when their native language lacks phonemic contrasts that occur in the target language. When L2 speakers experience problems being understood by voice-AI, they may naturally attempt to modify their language to be better understood by it, and this provides a natural setting for the production of clear speech.

Thus, the purpose of this study is to provide a systematic acoustic analysis of speech adaptions of L2 learners in interactions with voice-AI. We focused on the interaction between Amazon's Alexa and adult Korean L2 learners of English using English words containing vowel contrasts that are not found in their native language: /i/–/ɪ/ (as in *beat* vs. *bit*) and /ɛ/–/æ/ (as in *bet* vs. *bat*). We addressed two specific research questions. First, do L2 learners of English, compared to native controls, have difficulty being understood by voice-AI when producing English vowel contrasts that are non-existent in their native phonemic inventory? We tested this in a situation where Alexa is not able to determine the probability of the target word based on contextual clues. Second, when L2 learners are not understood in the initial attempt, do they make acoustic modifications with qualities of clear speech and target-like pronunciation in subsequent attempts?

Several important phonetic models, including the Revised Speech Learning Model (SLM-r) (Flege and Bohn, 2021), the Second Language Linguistic Perception Model (L2LP) (Escudero, 2005; Van Leussen and Escudero, 2015; Elvin and Escudero, 2019), and the Perception Assimilation Model (PAM) (Best, 1995) and PAM-L2 (Tyler et al., 2014), provide explanations of why L2 learners experience difficulty acquiring phonemic contrasts that are not in their native language. Based on the principles put forth by these models, we hypothesized that L2 learners would have difficulty being understood by voice-AI compared to native controls. Also, based on studies suggesting that native speakers make clear speech adaptations following misrecognition feedback from Alexa (Cohn and Zellou, 2021), we hypothesized that L2 participants will also make modifications to their speech when their initial production was not correctly recognized. The predicted modifications include clear speech and target-like pronunciation—that is, more similar to a model of native pronunciation.

As was mentioned earlier, this study focused on the production of /i/ and /ɪ/ (as in *beat* vs. *bit*), on the one hand, and /ɛ/ and /æ/ (as in *bet* vs. *bat*) on the other. In American English, these vowel pairs differ in their relative duration and spectral properties. The duration of /i/ is longer than /ɪ/, and /æ/ is longer than /ɛ/ (Flege et al., 1997). Compared to lax /ɪ/, tense /i/ occupies a higher and more anterior position in the vowel space, and therefore tends to have a lower F1 and a higher F2. Studies have shown that the spectral properties of /ɛ/ and /æ/ are highly variable among individual speakers of American English, resulting in a considerable degree of overlap in F1–F2 space (Hillenbrand et al., 1995). Because of the high degree of overlap in formants, duration is the primary cue in the /ɛ/–/æ/ contrast for native speakers of English (Hillenbrand et al., 2000). However, because /i/ and /ɪ/ are already sufficiently well-separated on the basis of spectral properties, duration plays only a small role in the recognition of this contrast (Hillenbrand et al., 2000).

Standard Korean is considered to have seven monophthongs /i, ɛ, ɯ, u, ʌ, ɑ, o/ (Shin, 2015). Due to Standard Korean lacking /i/–/ɪ/ and /ɛ/–/æ/ contrasts, Korean speakers experience difficulties in the production and perception of these English contrasts, which have been extensively studied (Flege et al., 1997; Tsukada et al., 2005; Baker and Trofimovich, 2006; Baker et al., 2008; Kim et al., 2018; Song and Eckman, 2019). For example, Flege et al. (1997) found that Korean speakers, both inexperienced and experienced with English, failed to produce significant duration and spectral differences between both /i/–/ɪ/ and /ɛ/–/æ/. An exception was inexperienced Korean participants who produced duration differences between /i/–/ɪ/.

To examine the interaction between Alexa and L2 learners, we conducted a speech production experiment in which Korean speakers and English-speaking controls asked Alexa

to spell out English words one at a time, using a fixed, context-free sentence structure. If L2 learners did not produce the target vowel, prompting a misrecognition response from Alexa, they were asked to attempt the same sentence again. We predicted decreased rates of accurate recognition by Alexa for Korean L2 learners, compared to controls. We also predicted that productions elicited after a misrecognition response would exhibit qualities of clear speech and target-like pronunciation.

Methods

Participants

We collected data from 20 participants. Half of them (5 female, 4 male, 1 preferred not to indicate) were monolingual, native-speakers of American English, serving as controls (*age range:* 22–42; $M = 30$). The other half (4 female, 6 male) were native speakers of Korean learning English as a second language (*age range:* 19–39; $M = 28$). The participants were recruited through campus advertisements at the University of Wisconsin-Milwaukee in the United States.

All participants in the native control group reported that they primarily use the North dialect (typically used in Wisconsin, Illinois, Minnesota, Michigan, etc.) of American English (Labov et al., 2008). Six participants reported some experience (beginner to intermediate) in learning a second language. One of the six participants reported intermediate knowledge of Korean.

All of our Korean participants had a relatively homogeneous experience with the target language: all of them started learning English as a second language in South Korea (beginning age ranged from 5 to 13 years, $M = 9$), have resided in the USA for 5 years or less, and use English for everyday study at an American university. We did not collect standardized proficiency scores from the participants, but Korean participants were asked to self-rate their English proficiency on the following scale: low beginner, high beginner, low intermediate, high intermediate, advanced. No pattern was found in this data, suggesting that participants had varying perceptions of their own proficiency level (see Appendix D in Supplementary material for the self-rated English proficiency provided by individual Korean participants).

None of the participants reported any speech or hearing problems, except for one native control, who reported a childhood stuttering problem. We also asked the participants if they had ever used an intelligent voice assistant before. Thirteen out of 20 indicated they had experience with one or more of the following intelligent voice assistants: Amazon's Alexa, Google's Google Assistant, Apple's Siri. Of these participants, four had experience with Amazon's Alexa, the voice-AI device we used in the present study.

Stimuli

As targets, we selected 64 monosyllabic CVC words. Half of these words formed minimal pairs for /i/ vs. /ɪ/, such as *seek* /sik/ and *sick* /sɪk/. The other half formed minimal pairs for /ɛ/ vs. /æ/, such as *pen* /pɛn/ and *pan* /pæn/. Our pilot experiment allowed us to develop exclusionary criteria for the target stimuli. First, we excluded any potential homophones, such as *peek/peak*. This was done in order to avoid the possibility that Alexa spelled the homophone instead of the target word. Second, during piloting, final voiced consonants were often partially devoiced and Alexa sometimes mis-interpreted them as voiceless. Therefore, we excluded words ending in a voiced consonant (e.g., *pig*) that has a minimal pair for coda voicing (e.g., *pick*). Third, since Korean does not have a contrast between /ɹ/ and /l/, and because they are difficult to separate from adjacent vowels, we also excluded words containing either of these sounds. In selecting fillers, we applied the same exclusionary criteria as for targets and selected 48 monosyllabic CVC words with vowels other than /i/, /ɪ/, /æ/, or /ɛ/, such as *move* /muv/. Filler words were not included in the analysis (see Appendicies A,B in Supplementary material for the full list of target and filler words).

To ensure that the Korean participants were familiar with the stimuli, we asked them to indicate any words that they did not know the meaning of. Nine out of 64 target words were marked unfamiliar by one or more participants. There was one word (*hem*) that was indicated as unfamiliar by 6 out of 10 participants; six words (*dim, deed, teak, den, gem, peck*) were indicated by two of the participants; two words (*heap, tan*) were indicated by one participant. To examine the frequency of these words, we referred to the Corpus of Contemporary American English (COCA), which contains more than one billion words of text from various sources such as spoken speech and written texts (https://www.english-corpora.org/coca/). Except for one word (*deed,* frequency = 10,479), all of the words indicated as unfamiliar have occurrence frequencies of 10,000 or less. Four words (*heap, dim, den, gem*) ranged from 10,000 to 5,000; three (*peck, hem, tan*) ranged from 5,000 to 1,000; one (*teak*) has a frequency of 793 occurrences. There were eight other words (*peach, knit, tick, mat, mash, jam, ham, mesh*) with frequencies under 10,000 in our stimulus list. These words were familiar to all participants. For the nine words that were marked unfamiliar by one or more participants, we allowed them to look up the words and the research assistant explained the words to their satisfaction.

Procedure

The experiment took place in a quiet laboratory setting, under the supervision of a research assistant. During the task, participants were seated next to an Alexa device (2nd Generation of Echo Show). Participants were provided with a printed list of

words, as well as a printed set of instructions that described the process for moving from one target word to the next. They were asked to produce each word within the frame sentence, "*Echo, spell* _____." To avoid potential effects of co-articulation, the instructions asked participants to pause between *spell* and the target word. Note that Alexa responds to one of four wake-up words, *Alexa, Amazon, Computer* or *Echo;* we chose *Echo* because it has fewer syllables and therefore takes less time to pronounce. Participants were instructed to monitor Alexa's response to each utterance, and were permitted as much time as they needed to process the response. If Alexa responded to the frame sentence by spelling the target word correctly (e.g., *S-E-E-K* for target word *seek*), the participants moved to the next trial. If Alexa responded with a different spelling (e.g., *S-I-C-K*), the participant produced the frame sentence a second time and, if necessary, a third time. After three productions, participants moved on to the next word in the list. The list of 112 stimulus words (64 targets + 48 fillers) was organized into two blocks, such that no two words from a minimal pair occurred in the same block. The order of the two blocks, as well as the order of the words within each block, was randomized for each participant. The productions were recorded directly onto a desktop computer running Audacity software at a sampling frequency of 44.1 kHz and 32-bit quantization *via* a Behringer XM8500 cardioid microphone that was located about five inches from the lips and connected to an M-Audio DMP3 preamplifier.

Analysis

We analyzed production data focusing on two aspects: Alexa responses for target, and acoustic properties of the vowel. For the Alexa responses, we coded the target, attempt number, and actual word Alexa spelled. Based on the Alexa responses, we had five categories, as shown in Table 1. When Alexa's response was the same as the target word or a homophone of the target word, the target word was judged to be correctly recognized by Alexa (although we had made every effort to eliminate homophones from the stimulus set, Alexa nevertheless occasionally gave unexpected homophone responses, such as the proper name *M-A-T-T* for the target word *mat*). In contrast, when Alexa's response to a target word was the other word of a minimal pair, or contained the other vowel of a minimal pair within a different word (e.g., *M-A-T-H* for target *met*), the target word was judged to be incorrectly recognized by Alexa. When Alexa's response to a target word was neither word of a minimal pair, we excluded the token, as the cause of the misrecognition was unclear. For example, if a word like *met* was recognized as *net*, we left out the data rather than counting it as an example of correct vowel recognition. Finally, if Alexa did not respond by spelling a word, the data was also eliminated.

Participants' vowel acoustic properties were analyzed in Praat (Boersma and Weenink, 2018). The beginning of each vowel was marked at the onset of F2, and the end was marked at the cessation of F2. For vowels bordered by a nasal, these boundaries were marked by abrupt decreases in amplitude and/or weakening in formant structure. For each vowel, we used a Praat script to automatically measure duration, F1 at midpoint, and F2 at midpoint. The default settings in Praat were used for formant tracking, with the maximum formant of 5,500 Hz for female speakers and 5,000 Hz for male speakers. Praat formant settings consistent with male speakers were used for the participant who preferred not to indicate their gender. If the formant tracking in Praat did not match the actual formant bands seen in the spectrogram, various manual adjustments were made to improve the formant tracking, such as adjusting the number of formants counted by Praat.

Results

We conducted two analyses, each addressing one of the two research questions. The aim of Analysis 1 was to examine whether Korean speakers, compared to native English controls, have difficulty being understood by Alexa when producing the English vowel contrasts that are non-existent in the Korean language. Figure 1 shows percent correct recognition of the four vowels (see Appendicies C,D in Supplementary material for the results for individual speakers). From the figure, it can be observed that Alexa recognized native controls' vowels more accurately than those of L2 speakers. To confirm the observation, we conducted two logit mixed-effects models, one for the /i/–/ɪ/ contrast and one for the /ɛ/–/æ/ contrast. Each model included the group (native controls vs. L2 speakers) and vowels (/i/ vs. /ɪ/ or /ɛ/ vs. /æ/) as fixed factors, and random intercepts for participants and words. The dependent variable was Alexa's recognition of each vowel token (correctly or incorrectly recognized). Statistical analyses were carried out in R using the *glmer* function in the *lmer4* package (Bates et al., 2015; R Development Core Team, 2016). *P*-values were estimated using the *lmerTest* package (Kuznetsova et al., 2017).

For /i/–/ɪ/, the result showed the significant effect of group ($z = -6.328$, $p < 0.001$), with higher recognition rate for native controls than L2 speakers. The effect of vowel ($z = -0.241$, $p = 0.81$) and the interaction between group and vowel ($z = 0.836$, $p = 0.403$) were not significant. For /ɛ/–/æ/, the effect of group was significant ($z = -10.124$, $p < 0.001$), with higher recognition rate for native controls than L2 speakers. Also, there was a significant interaction between group and vowel ($z = 2.07$, $p < 0.05$), suggesting that the difference between native controls and L2 speakers was larger for /æ/ compared to /ɛ/. There was no significant effect of vowel ($z = 1.073$, $p = 0.283$). In sum, for both vowel contrasts, it was found that L2 speakers are less accurately recognized by Alexa compared to native English controls.

The aim of Analysis 2 was to examine whether L2 speakers make acoustic modifications with qualities of clear speech and

TABLE 1 Number of tokens analyzed for each category.

Category	Judgment	Controls	L2 speakers	Total
When Alexa response is the same as the target word (e.g., mat → mat)	Correct	637	473	1,110
When Alexa response is a homophone of the target word (e.g., mat → Matt)	Correct	9	10	19
When Alexa response is the other word of a minimal pair (e.g., met → mat)	Incorrect	18	452	470
When Alexa response contains the other vowel of a minimal pair within a different word (e.g., met → math)	Incorrect	0	52	52
When Alexa response is neither word of a minimal pair (e.g., met → net, feet → faith)	Excluded	26	129	155
When Alexa doesn't spell a word (e.g., met → no response)	Excluded	6	139	145
	Total	696	1255	1,951

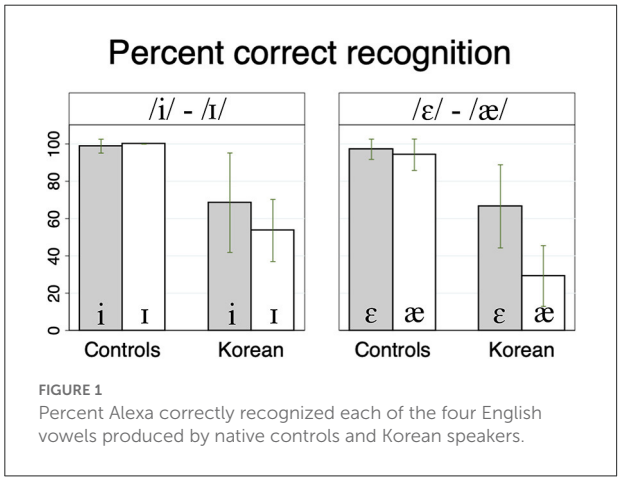

FIGURE 1
Percent Alexa correctly recognized each of the four English vowels produced by native controls and Korean speakers.

target-like pronunciation when they are not recognized by Alexa in the first attempt. Because most of the native controls' vowels were correctly recognized in the first attempt, thereby not providing a large enough number of tokens for comparison between the initial and subsequent attempts, Analysis 2 included only L2 speakers. We conducted Analysis 2 in two steps. First, we compared the acoustic properties of vowels correctly and incorrectly recognized by Alexa, to establish the acoustic properties of vowels recognized as target-like. Second, we compared the acoustic properties of vowels produced in the first attempts and in the subsequent attempts, to determine what kind of acoustic adjustments L2 learners make when misrecognition occurred in the first attempts.

In the first step, the mixed-effects regression models included correctness (correct or incorrect) as a fixed factor. In the second step, the mixed-effects regression models included attempts (initial or subsequent) as a fixed factor. Here, initial means a first attempt when L2 speakers produced a given word; subsequent attempts were second and third attempts in response to misrecognition. We combined the second and third attempts into one category, subsequent attempts, because our analysis found no systematic differences between the two. The dependent variable was one of five relevant acoustic properties: duration, F1, F2, vowel duration difference between contrasting vowels, and spectral distance between contrasting vowels. To calculate the spectral distances between two vowels, for example, between /i/ and /ɪ/, a general mathematical formula to calculate the Euclidean distance between two coordinates was used: $\{(F1_i\text{-}F1_I)^2 + (F2_i\text{-}F2_I)^2\}^{1/2}$. In the formula, "$F1_i$" and "$F2_i$" represent the F1 ($x$ coordinate) and F2 (y coordinate) for the vowel /i/, respectively. The distance between /ɛ/ and /æ/ was calculated in the same way.

The first three measures (duration, F1, F2) examined whether the acoustic properties of the four individual vowels, /i/, /ɪ/, /ɛ/, /æ/, changed as a function of correctness and attempts, whereas the next two measures (vowel duration difference, spectral distance) examined whether the acoustic differences between /i/–/ɪ/ on one hand, and /ɛ/–/æ/ on the other, changed as a function of correctness and attempts. Thus, for the first three measures, mixed-effect regression analyses were performed separately on each of the four vowels, /i/, /ɪ/, /ɛ/, /æ/, and for the next two measures, mixed-effect regression analyses were performed separately on each of the two vowel contrasts. Another difference was that for the first three measures, both participants and words were included as random factors, whereas for the next two measures, only participants were included as a random factor. In order to calculate the difference between contrasting vowels, we averaged duration and formant values for each vowel produced by individual participants. These averages were across words. Thus, it was not relevant to include word as a random factor to account for the differences between words. All mixed-effects regression analyses were carried out in R using the *lmer* function in the *lmer4* package, and p-values were estimated using the *lmerTest* package. The statistical results (t-values and their significance) from the analyses are presented in Table 2 (the effects of correctness) and Table 3 (the effects of attempts).

Figure 2 shows an overview of the number of tokens analyzed in both steps. There were 323 correctly recognized tokens and 235 incorrectly recognized tokens in the initial attempts. In the subsequent attempts, there were 160 correctly recognized tokens and 269 incorrectly recognized tokens. In the first step, where we examine the effect of correctness, 483 (323 initial + 160 subsequent) correct tokens were analyzed,

TABLE 2 Acoustic differences between vowels correctly and incorrectly recognized by Alexa (significant results are bolded).

	Vowel	Incorrect		Correct		Statistics
		Mean	*SE*	*Mean*	*SE*	
Duration (msec)	/i/	136	16	172	15.4	**$t = 5.229, p < 0.001$**
	/I/	122	9.4	103	9.2	**$t = -4.746, p < 0.001$**
	/ɛ/	176	12.8	147	12.4	**$t = -6.09, p < 0.001$**
	/æ/	165	12.4	188	13	**$t = 3.64, p < 0.001$**
F1 (Hz)	/i/	391	12.3	345	11.6	**$t = -7.167, p < 0.001$**
	/I/	355	13.1	420	12.4	**$t = 6.296, p < 0.001$**
	/ɛ/	781	25.5	766	23.1	$t = -0.878, p = 0.381$
	/æ/	766	30	850	31.4	**$t = 5.674, p < 0.001$**
F2 (Hz)	/i/	2,385	77.8	2,580	73.9	**$t = 5.069, p < 0.001$**
	/I/	2,501	61.9	2,340	59.5	**$t = -3.898, p < 0.001$**
	/ɛ/	1,903	64.8	1,902	60.8	$t = -0.035, p = 0.972$
	/æ/	1,927	54.5	1,861	57.6	**$t = -2.169, p < 0.05$**
Duration difference (msec)	/i/–/I/	−22.1	12.7	80.5	12.7	**$t = 6.246, p < 0.001$**
	/æ/–/ɛ/	−17.5	11	47.3	11	**$t = 4.167, p < 0.001$**
Distance (Hz^2)	/i/–/I/	199	41.9	286	41.9	$t = 1.933, p = 0.085$
	/æ/–/ɛ/	87	19.3	114	19.3	$t = 1.336, p = 0.218$

TABLE 3 Acoustic differences between vowels produced in the subsequent attempts in response to a misrecognition by Alexa in the first attempts (significant results are bolded).

	Vowel	Initial attempts		Subsequent attempts		Statistics
		Mean	*SE*	*Mean*	*SE*	
Duration (msec)	/i/	133	13.9	169	13.3	**$t = 5.625, p < 0.001$**
	/I/	127	10.5	109	10.1	**$t = -4.178, p < 0.001$**
	/ɛ/	172	15.3	164	14.6	$t = -1.163, p = 0.248$
	/æ/	162	12.6	177	12.2	**$t = 2.644, p < 0.01$**
F1 (Hz)	/i/	381	14	356	12.9	**$t = -3.04, p < 0.01$**
	/I/	361	15.8	396	15.1	**$t = 4.346, p < 0.001$**
	/ɛ/	779	30.1	772	26.6	$t = -0.352, p = 0.726$
	/æ/	757	34.3	793	33.6	**$t = 2.872, p < 0.01$**
F2 (Hz)	/i/	2,385	78.1	2,492	73.2	**$t = 2.539, p < 0.05$**
	/I/	2,518	70.1	2,391	63.9	**$t = -2.529, p < 0.05$**
	/ɛ/	1,882	71.2	1,927	65.8	$t = 1.122, p = 0.265$
	/æ/	1,954	59.4	1,897	57.6	**$t = -2.062, p < 0.05$**
Duration difference (msec)	/i/–/I/	−27.5	16.7	53.9	16.7	**$t = 5.556, p < 0.001$**
	/æ/–/ɛ/	−20.99	10.2	−4.19	10.2	**$t = 2.577, p < 0.05$**
Distance (Hz^2)	/i/–/I/	213	31.4	135	31.4	$t = -1.762, p = 0.095$
	/æ/–/ɛ/	92.3	18.1	84.8	18.1	$t = -0.303, p = 0.771$

compared to 504 (235 initial + 269 subsequent) incorrect. In the second step, where we examine the effect of attempts following an initial misrecognition, we only included initial incorrect tokens (235), which was compared against all 429 (160 correct + 269 incorrect) subsequent attempts.

First, we found that the vowels correctly and incorrectly recognized by Alexa differed in most acoustic measures we employed, and that the differences are in the expected direction. As shown in Table 2, the durations of /i/ and /æ/ were significantly longer, and the durations of /ɪ/ and /ɛ/ were significantly shorter when they were correctly recognized compared to misrecognized. F1 and F2 also differed significantly for all vowels, except for /ɛ/. For correctly recognized /i/, F1 was lower and F2 was higher than incorrectly recognized /i/. For

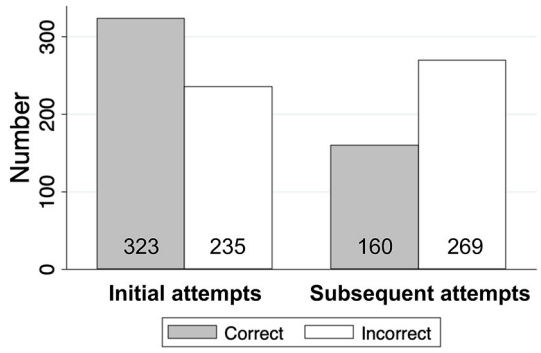

FIGURE 2
The number of tokens analyzed in Analysis 2.

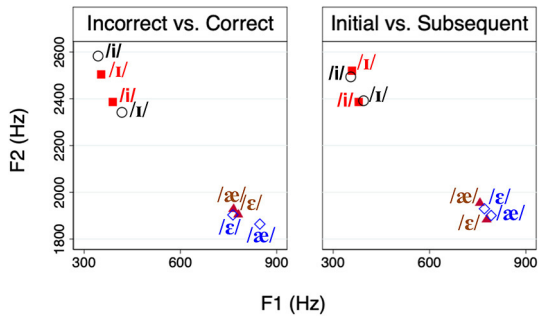

FIGURE 3
Contrasting vowels on the F1–F2 plane. Filled markers show the incorrect(left)/initial(right) condition, whereas hollow markers show the correct(left)/subsequent(right) condition.

/ɪ/ and /æ/, which occupy a lower and a more back position than their respective counterparts, /i/ and /ɛ/, F1 was higher and F2 was lower when correctly recognized. For both vowel contrasts, the difference in duration between vowels was larger when correctly recognized. As can be seen in both Table 2 and Figure 3, spectral distance between contrasting vowels overall increased when they were correctly recognized by Alexa, although the difference was not found to be significant in either pair.

Second, we compared the acoustic properties of vowels produced in the first attempts and in the subsequent attempts to examine modifications following a misrecognition by Alexa in the first attempts. For this reason, the initial attempts in Table 3 included only those incorrectly recognized in the first attempts. Subsequent attempts included both correctly and incorrectly recognized second and third attempts. We found a striking similarity between the results presented in Tables 2, 3. As shown in Table 3, the properties of subsequent attempts were similar to correctly recognized vowels in Table 2. This suggested that when L2 learners are not understood in the initial attempt, they made acoustic modifications with qualities of target-like pronunciation in the subsequent attempts. For three of the vowels, /i/, /ɪ/, and /æ/, duration, F1, and F2 in the subsequent attempts changed in the direction expected for the vowels. For example, the durations of /i/ and /æ/ were significantly longer in the subsequent attempts than the initial attempt, whereas the duration of /ɪ/ was significantly shorter in the subsequent attempts. However, /ɛ/ showed no change in the three measures between the initial and the subsequent attempts. For both vowel contrasts, the difference in duration between vowels was larger in the later attempts compared to the first attempts. Just like no increase in spectral distance was found between incorrectly and correctly recognized vowels, we found no increase in spectral distance between the initial and subsequent attempts. Although F1 and F2 of individual vowels changed in the expected direction (except for /ɛ/) (see Figure 3), these changes did not necessarily increase the distance between contrasting vowels significantly. This suggested that increasing spectral distance between contrasting vowels was not a way L2 speakers responded to misrecognitions of initial attempts.

Discussion

The current study offers two key findings. First, Amazon's Alexa provided significantly lower rates of target recognition for individual words produced by Korean L2 speakers of English, compared to native English-speaking controls. Notably, the patterns of recognition differed from one vowel to the next. Second, in response to a misrecognition by Alexa, L2 speakers adjusted subsequent productions of most vowels, such that they differed acoustically from initial productions. These adjustments exhibited some, but not all, of the predicted characteristics of clear speech and target-like pronunciation. We discuss each of these findings in turn.

Alexa's performance in target word recognition differed substantially across the two speaker groups: while its overall recognition rate for native English productions was 98%, its rate for L2 productions was 55%. Of course, these figures are certainly not representative of Alexa's interactions with L2 speakers in general, because they are restricted to sentences without semantic context ("Echo, spell _____.") and because our target words contained vowel contrasts that are known to be difficult for Korean L2 speakers. Nevertheless, the specific patterning of these results can help us to better characterize the nature of L2 vowel pronunciations. For example, while Alexa's recognition rates for L2 productions of words with /ɛ/ were relatively high (67%), rates for words with /æ/ were noticeably low (29%), and our statistical analysis revealed a significant interaction between these two vowels and speaker groups (L2 vs. control). In a similar vein, the descriptive results suggest that recognition rates for words with /i/ were relatively high (68%) while rates for words with /ɪ/ were essentially at chance (54%), although the statistical analysis did not reveal a significant

TABLE 4 Comparison of recognition rates by Alexa and human listeners.

	/i/	/ɪ/	/ɛ/	/æ/
Alexa recognition rate (current study)	68	54	67	29
Human recognition rate (Song and Eckman, under review)	66	53	82	39

interaction in this case. These asymmetries clearly skew in favor of those vowels, /ɛ/ and /i/, that occur in the speakers' native L1 inventory. Even for these vowels, however, Alexa's recognition rates for L2 speakers were still lower than those for native English controls (99% for /i/, 97% for /ɛ/) suggesting that L2 speakers' production targets retain vestigial characteristics of their L1.

The different rates of recognition for /ɪ/ compared to /æ/ are also worth noting. Alexa's chance-level recognition for L2 productions with /ɪ/ (54%) suggests that the speakers have achieved at least partial success at forming a new L2 vowel category that is distinct from /i/. By contrast, Alexa's very low recognition rate for L2 productions with /æ/ (29%) suggests that speakers have not yet succeeded in forming a new L2 vowel category, and overwhelmingly produce /æ/ targets as /ɛ/ instead. Recall that our L2 participant pool exhibited very high levels of spoken English language ability overall, and were operating successfully in an English-speaking university environment. Thus, our results are consistent with earlier findings (Tahta et al., 1981) demonstrating that, despite functional fluency in a second language, subtle but significant pronunciation difficulties may persist.

As noted in the Introduction, several previous studies have focused on whether voice-AI is useful as a pedagogical tool for learning L2 pronunciation. In particular, these studies have noted that L2 speakers may be more relaxed and engaged with an assistant such as Alexa, compared to a human instructor. In order for voice-AI devices to be truly useful in this regard, however, their responses to L2 productions should be similar to human responses. In other words, we want Alexa to "hear" the same thing that humans hear. Our results suggest that, at least for the L2 vowel contrasts examined here, this is indeed the case. Table 4 displays Alexa recognition rates for L2 productions from the current study, compared with human recognition rates for similar L2 productions from Song and Eckman (under review).

In the Song and Eckman (under review) study, native speakers of English listened to Korean L2 single-word CVC productions similar to those examined here, such as *seek, sick, set,* and *sat*. The listeners transcribed what they heard, which was compared to the L2 speakers' intended targets. As is evident in Table 4, the human recognition rates pattern quite similarly to Alexa recognition rates. In both cases, rates are overall higher for /i/ and /ɛ/, the English vowels that also occur in the Korean inventory. And in both cases, rates are essentially at chance for /ɪ/, and notably low for /æ/. Results of unpaired *t*-tests revealed no significant differences between recognition rates across experiments, for any of the four vowels (for /i/, $p = 0.82$; for /ɪ/, $p = 0.94$; for /ɛ/, $p = 0.12$; for /æ/, $p = 0.31$). The striking similarity in human versus Alexa recognition rates suggests that, in the future, voice-AI could indeed serve as an effective diagnostic tool for L2 pronunciation.

Turning to the question of misrecognition, our results showed that, when making a second or third attempt to get Alexa to spell the target word, L2 speakers adjusted their productions such that they were acoustically different from their initial attempt. That is, in response to Alexa's misrecognition, L2 speakers' vowel duration, F1, and F2 changed in directions that strengthen the vowels' phonological features, suggesting clear speech and target-like pronunciation. For /i/ and /æ/, the vowels that are intrinsically longer in duration and occupy more peripheral positions in the vowel space than their respective counterparts, /ɪ/ and /ɛ/, later attempts were significantly longer in duration than the initial attempt. In addition, later attempts of /i/ exhibited lower F1 and higher F2, while later attempts of /æ/ exhibited higher F1 and lower F2. Meanwhile, for the vowel /ɪ/, later attempts were significantly shorter in duration, higher in F1, and lower in F2 than the initial attempt. The only vowel which exhibited no change was /ɛ/, whose duration, F1, and F2 differences between initial and later attempts were not significant. Another key finding of the present study is that the differences in duration between /i/–/ɪ/, on the one hand, and /ɛ/–/æ/ on the other, increased in the later attempts compared to the initial attempt. Overall, these patterns are consistent with the prediction that, in response to a misrecognition by Alexa, L2 speakers attempted to produce clear speech. With that being said, L2 speakers did not respond to misrecognitions by increasing spectral distance between contrasting vowels. Thus, although L2 speakers did make adjustments to F1 and F2 of three of the four vowels following a misrecognition, these adjustments did not necessarily increase the distance between contrasting vowels significantly. The robust and consistent adjustments in vowel duration are consistent with previous work examining the productions of Korean learners of English, which have shown that they tend to implement the /i/–/ɪ/ and /ɛ/–/æ/ contrasts primarily through duration differences, rather than formant differences (Flege et al., 1997; Kim et al., 2018).

The ultimate goal of pronunciation training is accurate recognition by the listener, which can be achieved with the production of native-like targets. However, in our study paradigm, when Alexa did not recognize a target, the L2 participants did not adjust their productions so as to produce more target-like /ɛ/. Also, recognition rate of /æ/ was very low, only at 29%. It is possible that Alexa's response (e.g., Saying "*Pat* is spelled *P-A-T*" in response to the intended target *pet*) did not provide appropriate or sufficient information. For example, the

response does not contain the target vowel /ɛ/, and therefore did not provide L2 speakers with any pronunciation target, and the same point can be made for Alexa's responses to utterances with other vowels. Furthermore, as noted earlier, our analysis indicated no significant acoustic differences between vowels produced on second vs. third attempts, suggesting that there may be a ceiling effect on the degree of L2 speaker adjustments. Future work could address some of these issues by asking L2 participants to engage in more varied tasks with Alexa that expose them to the pronunciation of both the target sound and the actual sound Alexa heard.

In our review of the literature on interactions between voice-AI and L2 learners, we saw that the recognition rates reported by previous studies vary quite widely (Dizon, 2017; Moussalli and Cardoso, 2020; Dizon et al., 2022). The current study suggests that we might better understand these patterns if we can pinpoint the exact features of L2 speech that give rise to communication breakdowns in the first place. For example, our results show that the difference between native controls and L2 speakers was larger for /æ/ compared to /ɛ/. This suggests that words which contain sounds not found in the L2 inventory will be more susceptible to misrecognition. Future work could broaden this hypothesis beyond the four vowel phonemes tested here. In addition, several previous studies had noted that, when voice-AI does not recognize L2 utterances, learners tend to abandon their communication attempt (Dizon and Tang, 2020; Dizon et al., 2022; Tai and Chen, 2022). The current study suggests a potential paradigm for encouraging learners to repeat their utterance, rather than abandoning it, and thereby to encourage pronunciation adjustments. Future work could refine this paradigm, for example, by having Alexa produce correct targets as pronunciation models.

Overall, our findings show that recognition rates are low for L2 learners in a situation where they interact with voice-AI using the L2 vowel contrast that does not occur in their native language, and where they receive no information about what the correct or wrong pronunciation is like. Nonetheless, presence of misrecognition itself directed L2 participants' attention toward errors in their pronunciation, thereby leading to more target-like productions in the subsequent attempts. Thus, this study provides strong evidence that even in such an adverse situation, L2 participants make various clear speech modifications in response to misrecognition by voice-AI.

Conclusion

The current study examined how Amazon's Alexa responded to Korean L2 learners of English, focusing on words with one of two vowel contrasts, /i/ vs. /ɪ/ and /æ/ vs. /ɛ/, which do not occur in Korean. Our results showed that Alexa was less accurate at identifying utterances produced by L2 learners, compared to those produced by controls, and that patterns of recognition differed from one vowel to the next. Our results also showed that, when Alexa misrecognized a word, L2 speakers adjusted subsequent productions of most vowels, exhibiting some features of clear speech. Although our study contained a relatively small number of participants and focused on a restricted set of vowel contrasts produced by members of one L2 population, future work can include larger numbers of participants, different types of speech sound contrasts, and additional L2 populations. This research platform can inform our understanding of voice-AI as a potentially powerful vehicle for language learning, as well as our understanding of speech production more generally.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving human participants were reviewed and approved by the Institutional Review Board (IRB) at the University of Wisconsin-Milwaukee. The patients/participants provided their written informed consent to participate in this study.

Author contributions

JS and AP contributed to conceptualization and design of the study. JS and TC ran experiments. TC coded the data. JS performed analyses. All authors contributed to manuscript drafting. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by University of Wisconsin-Milwaukee's Research Assistance Fund.

Acknowledgments

We thank Annika Huber and Marko Pavlovic for their research assistance.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships

that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcomm.2022.995475/full#supplementary-material

References

Baker, W., and Trofimovich, P. (2006). Perceptual paths to accurate production of L2 vowels: the role of individual differences. *IRAL* 44, 231–250. doi: 10.1515/IRAL.2006.010

Baker, W., Trofimovich, P., Flege, J. E., Mack, M., and Halter, R. (2008). Child-adult differences in second-language phonological learning: the role of cross-language similarity. *Lang. Speech.* 51, 317–342. doi: 10.1177/0023830908099068

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Soft.* 67, 1–48. doi: 10.18637/jss.v067.i01

Best, C. (1995). "A direct realist view of cross-language speech perception," in *Speech Perception and Linguistic Experience: Issues in Cross-Language Research*, ed W. Strange (Timonium, MD: New York Press), 171–204.

Boersma, P., and Weenink, D. (2018). *Praat: Doing Phonetics by Computer (Version 6.0.37) [Computer software].* Retrieved from http://www.praat.org/

Bradlow, A. R., and Bent, T. (2002). The clear speech effect for non-native listeners. *J. Acous. Soc. Am.* 112, 272–284. doi: 10.1121/1.1487837

Bradlow, A. R., Kraus, N., and Hayes, E. (2003). Speaking clearly for learning-impaired children: sentence perception in noise. *J. Speech Lang. Hear. Res.* 46, 80–97. doi: 10.1044/1092-4388(2003/007)

Bradlow, A. R., Torretta, G. M., and Pisoni, D. B. (1996). Intelligibility of normal speech I: global and fine-grained acoustic-phonetic talker characteristics. *Speech Commun.* 20, 255–272. doi: 10.1016/S0167-6393(96)00063-5

Brumm, H., and Zollinger, S. A. (2011). The evolution of the Lombard effect: 100 years of psychoacoustic research. *Behaviour* 148, 1173–1198. doi: 10.1163/000579511X605759

Burnham, D. K., Joeffry, S., and Rice, L. (2010). "Computer-and human-directed speech before and after correction," in *Proceedings of the 13th Australasian International Conference on Speech Science and Technology* (Melbourne, VIC), 13–17.

Chen, H. H. J., Yang, C. T. Y., and Lai, K. K. W. (2020). Investigating college EFL learners' perceptions toward the use of google assistant for foreign language learning. *Interact. Learn. Envir.* 1–16. doi: 10.1080/10494820.2020.1833043. [Epub ahead of print].

Cohn, M., Liang, K., Sarian, M., Zellou, G., and Yu, Z. (2021). Speech rate adjustments in conversations with an Amazon Alexa socialbot. *Front. Commun.* 6, 671429. doi: 10.3389/fcomm.2021.671429

Cohn, M., Segedin, B. F., and Zellou, G. (2022). The acoustic-phonetic properties of Siri- and human-directed speech. *J. Phon.* 90, 101123. doi: 10.1016/j.wocn.2021.101123

Cohn, M., and Zellou, G. (2021). Prosodic differences in human- and Alexa-directed speech, but similar error correction strategies. *Front. Commun.* 6, 675704. doi: 10.3389/fcomm.2021.675704

Dizon, G. (2017). Using intelligent personal assistants for second language learning: a case study of Alexa. *TESOL J.* 8, 811–830. doi: 10.1002/tesj.353

Dizon, G. (2020). Evaluating intelligent personal assistants for L2 listening and speaking development. *Lang. Learn. Technol.* 24, 16–26.

Dizon, G., and Tang, D. (2020). Intelligent personal assistants for autonomous second language learning: an investigation of Alexa. *JALT CALL J.* 16, 107–120.

Dizon, G., Tang, D., and Yamamoto, Y. (2022). A case study of using Alexa for out-of-class, self-directed Japanese language learning. *Comput. Educ. Artif. Intell.* 3, 100088. doi: 10.1016/j.caeai.2022.100088

Elvin, J., and Escudero, P. (2019). "Cross-linguistic influence in second language speech: implications for learning and teaching," in *Cross-Linguistic Influence: From Empirical Evidence to Classroom Practice*, eds J. Gutierrez-Mangado, M. Martínez-Adrián, and F. Gallardo-del-Puerto (Cham: Springer), 1–20. doi: 10.1007/978-3-030-22066-2_1

Escudero, P. (2005). *Linguistic perception and second language acquisition: Explaining the attainment of optimal phonological categorization.* (Ph.D. thesis), Utrecht University (Utrecht: Netherlands).

Ferguson, S. H. (2004). Talker differences in clear and conversational speech: vowel intelligibility for normal-hearing listeners. *J. Acoust. Soc. Am.* 116, 2365–2373. doi: 10.1121/1.1788730

Ferguson, S. H., and Kewley-Port, D. (2002). Vowel intelligibility in clear and conversational speech for normal-hearing and hearing-impaired listeners. *J. Acoust. Soc. Am.* 112, 259–271. doi: 10.1121/1.1482078

Ferguson, S. H., and Kewley-Port, D. (2007). Talker differences in clear and conversational speech: acoustic characteristics of vowels. *J. Speech Lang. Hear. R.* 50, 1241–1255. doi: 10.1044/1092-4388(2007/087)

Ferguson, S. H., and Quené, H. (2014). Acoustic correlates of vowel intelligibility in clear and conversational speech for young normal-hearing and elderly hearing-impaired listeners. *J. Acoust. Soc. Am.* 135, 3570–3584. doi: 10.1121/1.4874596

Flege, J. E., Bohn, O.-C., and Jang, S. (1997). The effect of experience on nonnative subjects' production and perception of English vowels. *J. Phon.* 25, 437–470. doi: 10.1006/jpho.1997.0052

Flege, J. E., and Bohn, O. S. (2021). "The revised speech learning model (SLM-r)," in *Second Language Speech Learning: Theoretical and Empirical Progress,* ed R. Wayland (Cambridge: Cambridge University Press), 3–83. doi: 10.1017/9781108886901.002

Hazan, V., Grynpas, J., and Baker, R. (2012). Is clear speech tailored to counter the effect of specific adverse listening conditions? *J. Acoust. Soc. Am.* 132, EL371–EL377. doi: 10.1121/1.4757698

Hillenbrand, J., Getty, L., Clark, M., and Wheeler, K. (1995). Acoustic characteristics of American English vowels. *J. Acoust. Soc. Am.* 97, 3099–3111. doi: 10.1121/1.411872

Hillenbrand, J. M., Clark, M. J., and Houde, R. A. (2000). Some effects of duration on vowel recognition. *J. Acoust. Soc. Am.* 108, 3013–3022. doi: 10.1121/1.1323463

Kim, D., Clayards, M., and Goad, H. (2018). A longitudinal study of individual differences in the acquisition of new vowel contrasts. *J. Phon.* 67, 1–20. doi: 10.1016/j.wocn.2017.11.003

Krause, J. C., and Braida, L. D. (2002). Investigating alternative forms of clear speech: the effects of speaking rate and speaking mode on intelligibility. *J. Acoust. Soc. Am.* 112, 2165–2172. doi: 10.1121/1.1509432

Krause, J. C., and Braida, L. D. (2004). Acoustic properties of naturally produced clear speech at normal speaking rates. *J. Acoust. Soc. Am.* 115, 362–378. doi: 10.1121/1.1635842

Kuhl, P. K., Andruski, J. E., Chistovich, I. A., Chistovich, L. A., Kozhevnikova, E. V., Ryskina, V. L., et al. (1997). Cross-language analysis of phonetic units in language addressed to infants. *Science* 277, 684–686. doi: 10.1126/science.277.5326.684

Kuznetsova, A., Brockhoff, P. B., and Christensen, R. H. (2017). lmerTest package: tests in linear mixed effects models. *J. Stat. Soft.* 82, 1–26. doi: 10.18637/jss.v082.i13

Labov, W., Ash, S., and Boberg, C. (2008). *The Atlas of North American English: Phonetics, Phonology and Sound Change*. Berlin: Walter de Gruyter.

Lindblom, B. (1990). "Explaining phonetic variation: a sketch of the H&H theory," in *Speech Production and Speech Modelling, Vol. 55*, eds W. J. Hardcastle, and A. Marchal (Dordrecht: Kluwer Academic Publishers), 403–439. doi: 10.1007/978-94-009-2037-8_16

Moussalli, S., and Cardoso, W. (2016). "Are commercial 'personal robots' ready for language learning? Focus on second language speech," in *CALL communities and culture–short papers from EUROCALL,* eds S. Papadima-Sophocleous, L. Bradley, and S. Thouësny (Dublin: Research-publishing.net), 325–329. doi: 10.14705/rpnet.2016.eurocall2016.583

Moussalli, S., and Cardoso, W. (2020). Intelligent personal assistants: can they understand and be understood by accented L2 learners? *Comput. Assist. Lang. Learn.* 33, 865–890. doi: 10.1080/09588221.2019.1595664

Picheny, M. A., Durlach, N. I., and Braida, L. D. (1985). Speaking clearly for the hard of hearing I: intelligibility differences between clear and conversational speech. *J. Speech Lang. Hear. Res.* 28, 96–103. doi: 10.1044/jshr.2801.96

Picheny, M. A., Durlach, N. I., and Braida, L. D. (1986). Speaking clearly for the hard of hearing II: acoustic characteristics of clear and conversational speech. *J. Speech Hear. Res.* 29, 434–446. doi: 10.1044/jshr.2904.434

R Development Core Team (2016). *R: A Language and Environm ent for Statistical Computing*. Vienna: R foundation for Statistical Computing (Version 3.6.3) [Computer software]. Retrieved from http://www.R-project.org.

Shin, J. (2015). "Vowels and consonants," in *The Handbook of Korean Linguistics*, eds L. Brown,and J. Yeon (Chichester: Wiley-Blackwell), 3–21. doi: 10.1002/9781118371008.ch1

Smiljanić, R., and Bradlow, A. R. (2005). Production and perception of clear speech in Croatian and English. *J. Acoust. Soc. Am.* 118, 1677–1688. doi: 10.1121/1.2000788

Smiljanić, R., and Bradlow, A. R. (2009). Speaking and hearing clearly: talker and listener factors in speaking style changes. *Lang. Ling. Compass* 3, 236–264. doi: 10.1111/j.1749-818X.2008.00112.x

Song, J. Y., Demuth, K., and Morgan, J. (2010). Effects of the acoustic properties of infant-directed speech on infant word recognition. *J. Acoust. Soc. Am.* 128, 389–400. doi: 10.1121/1.3419786

Song, J. Y., and Eckman, F. (2019). Covert contrasts in the acquisition of English high front vowels by native speakers of Korean, Portuguese, and Spanish. *Lang. Acquis.* 26, 436–456. doi: 10.1080/10489223.2019.1593415

Song, J. Y., and Eckman, F. (under review). The relationship between second-language learners' production and perception of English vowels: The role of target-like acoustic properties. *Second Lang. Res.*

Tahta, S., Wood, M., and Loewenthal, K. (1981). Foreign accents: factors relating to transfer of accent from the first language to a second language. *Lang. Speech.* 24, 265–272. doi: 10.1177/002383098102400306

Tai, T. Y., and Chen, H. H. J. (2022). The impact of intelligent personal assistants on adolescent EFL learners' listening comprehension. *Comput. Assist. Lang. Learn.* 1–28. doi: 10.1080/09588221.2022.2040536. [Epub ahead of print].

Tsukada, K., Birdsong, D., Bialystok, E., Mack, M., Sung, H., and Flege, J. (2005). A developmental study of English vowel production and perception by native Korean adults and children. *J. Phon.* 33, 263–290. doi: 10.1016/j.wocn.2004.10.002

Tyler, M., Best, C., Faber, A., and Levitt, A. (2014). Perceptual assimilation and discrimination of non-native vowel contrasts. *Phonetica* 71, 4–21. doi: 10.1159/000356237

Underwood, J. (2017). "Exploring AI language assistants with primary EFL students," *in CALL in a climate of change: adapting to turbulent global conditions–short papers from EUROCALL*, eds K. Borthwick, L. Bradley, and S. Thouësny (Dublin: Research-publishing.net), 317–321. doi: 10.14705/rpnet.2017.eurocall2017.733

Uther, M., Knoll, M. A., and Burnham, D. (2007). Do you speak E-NG-LI-SH? A comparison of foreigner-and infant-directed speech. *Speech Commun.* 49, 2–7. doi: 10.1016/j.specom.2006.10.003

Van Leussen, J. W., and Escudero, P. (2015). Learning to perceive and recognize a second language: the L2LP model revised. *Front. Psychol.* 6, 1000. doi: 10.3389/fpsyg.2015.01000