Korean Journal of English Language and Linguistics, **Vol 26, July 2026, pp. 808-832**
DOI: 10.15738/kjell.26..202607.808

KOREAN JOURNAL OF ENGLISH LANGUAGE AND LINGUISTICS

ISSN: 1598-1398 / e-ISSN 2586-7474

http://journal.kasell.or.kr

AI-Generated versus Human-Recorded Voice Input for L2 English Vowel Perception: A Classroom-Based Perception-Production Training with Korean EFL Middle School Learners

Hyoyoung Park · Hyunkee Ahn (Seoul National University)

Received: March 10, 2026
Revised: May 27, 2026
Accepted: July 3, 2026

Park, Hyoyoung (First author)
Ph.D. candidate, Department of English Language Education
Seoul National University
Email: hpark412@snu.ac.kr

Ahn, Hyunkee (Corresponding author)
Professor, Department of English Language Education
Seoul National University
1 Gwanak-ro, Gwanak-gu
Seoul, Korea
Tel: +82-2-880-7673
Email: ahnhk@snu.ac.kr

ABSTRACT

Park, Hyoyoung and Hyunkee Ahn. 2026. AI-generated versus human-recorded voice input for L2 English vowel perception: A classroom-based perception-production training with Korean EFL middle school learners. *Korean Journal of English Language and Linguistics 26*, 808-832.

This study examined whether AI-generated speech can effectively support classroom-based, high-variability phonetic training (HVPT)-informed pronunciation training that combines perception and production practice in L2 vowel learning, and whether its effects differ from those of human-recorded speech in terms of learning magnitude, generalization, and contrast-specific outcomes. Seventy-four Korean middle school learners of English completed 8 weeks of pronunciation training in which identification and production practice were delivered with either AI-generated or human-recorded stimuli targeting six English vowel contrasts. Phoneme identification accuracy was assessed using a two-alternative forced-choice (2AFC) task before and after training, with generalization examined across voice types (new vs. familiar), word types (trained vs. untrained), and vowel contrasts. Data were analyzed using binomial generalized linear mixed-effects models. Results showed significant pre-post improvement in both groups, with the AI-voice group demonstrating greater overall gains. Learners in both groups generalized to new voices and untrained lexical items. However, no significant three-way interactions were found for voice or word-type generalization, indicating that the pattern of transfer did not differ between training conditions. In addition, although improvement varied across vowel contrasts, the relative hierarchy of phonological difficulty was preserved across groups. These findings suggest that AI-generated speech can enhance the magnitude of phonetic learning without altering the underlying mechanisms of generalization or the organization of L2 vowel perception. Moreover, AI-generated speech appears to support category-level restructuring in ways comparable to human speech input. The greater gains observed in the AI condition may be related to differences in the variability and consistency of AI-generated input, a possibility that warrants further investigation. Overall, the results highlight the pedagogical potential of AI-generated speech as a scalable and theoretically grounded tool for L2 pronunciation training.

KEYWORDS

AI-generated speech, neural text-to-speech (TTS), pronunciation training, high variability phonetic training (HVPT), L2 vowel perception, generalization, Korean EFL learners

1. Introduction

The acquisition of second language (L2) phonological contrasts remains one of the most persistent challenges for L2 learners (Flege 1995, Flege and Bohn 2021). For Korean learners of English in particular, English vowel contrasts that do not map transparently onto the L1 vowel system often lead to long-lasting perceptual difficulties (Hong 2012). High-variability phonetic training (HVPT) has been shown to facilitate perceptual restructuring by exposing learners to multiple talkers and diverse acoustic realizations of target contrasts (Thomson 2012), thereby promoting category-level abstraction rather than item-specific memorization. Within the framework of the Speech Learning Model (SLM), such variability plays a crucial role in destabilizing L1-based perceptual mappings and supporting the formation or refinement of L2 phonetic categories (Flege and Bohn 2021, Thomson 2012).

Traditionally, HVPT relies on recordings produced by multiple native speakers. However, recent advances in speech synthesis technology have made it possible to generate highly naturalistic AI-based speech with controllable acoustic properties (Barrington et al. 2025, Bione and Cardoso 2020). This development raises an important theoretical and pedagogical question: Can AI-generated speech induce training effects comparable to those of human-recorded input in classroom-based pronunciation instruction? While AI-based materials are increasingly used in language learning contexts (Mohammadkarimi 2024, Vancova 2023), empirical evidence regarding their effectiveness in HVPT context remains limited. By distinguishing between learning strength and learning structure, this study aims to clarify not only whether AI-generated speech works, but how it works within established models of L2 phonological development and HVPT framework.

2. Literature Review

2.1 The Speech Learning Model (SLM and SLM-r)

The Speech Learning Model (SLM) proposes that L2 speech acquisition is strongly influenced by the perceived phonetic similarity between L1 and L2 sounds (Flege 1995). When learners perceive an L2 sound as similar to an existing L1 category, they tend to assimilate it into that existing category rather than form a new one, leading to persistent perceptual difficulty. In contrast, if learners find an L2 sound sufficiently distinctive, they are more likely to form a new category for it. The revised version of Speech Learning Model (SLM-r) further emphasizes that phonetic learning is a process that happens across the lifespan, as long as learners receive adequate input and language experience (Flege and Bohn 2021). From this perspective, successful speech learning involves the restructuring and refinement of existing L1-based perceptual mappings as learners gradually develop more accurate phonetic representations for L2 sounds (Bohn and Munro 2007). Thus, L2 speech learning can be considered in terms of both learning magnitude, reflected in improvements in perceptual accuracy, and learning process involving the restructuring of underlying phonetic categories.

Beyond this dual framework, SLM-r also emphasizes that perception and production are not independent processes but develop in tandem throughout L2 acquisition. According to the model, perceptual representations both guide and are shaped by production attempts: learners' improving perception of L2 sounds supports more accurate production, while production practice in turn refines perceptual categories (Flege and Bohn 2021, Nagle and Baese-Berk 2022). This bidirectional perception-production link suggests that training programs combining perceptual and production-based activities may engage these two processes in mutually reinforcing ways, potentially accelerating phonetic learning beyond what perception-only exposure would achieve.

Importantly, SLM-r highlights the role of input quality and variability in facilitating such perceptual restructuring. Exposure to varied and adequate L2 speech input allows learners to detect acoustic differences between similar L1 and L2 sounds and supports the formation of new phonetic categories. This issue is particularly relevant in English-as-a-foreign-language (EFL) contexts, where natural exposure to diverse L2 input is limited. Under these conditions, structured training that provides high-quality and variable input can play a critical role in promoting perceptual learning and phonetic category development (Piske 2007, Tyler 2019).

2.2 L2 Speech Perception and English Vowel Identification by Korean Learners

A substantial body of previous research has documented persistent difficulties among Korean L2 learners in distinguishing English vowel contrasts, often to a greater extent than consonants (Hong 2012, Park and Lee 2024, Yang 2010). One major source of difficulty lies in the structural differences between the Korean and English vowel systems and consequent perceptual differences. Korean has a relatively smaller and more "symmetrical" vowel inventory, whereas English contains a larger set of vowel categories with finer phonetic distinctions (Kim 2010, Lee and Iverson 2012, p. 1452). As a result, multiple English vowels may be perceptually mapped onto a single Korean category, making certain English contrasts particularly difficult for Korean learners to discriminate. While these mapping patterns are broadly consistent with the predictions of the Perceptual Assimilation Model for L2 (PAM-L2; Best and Tyler 2007), Korean listeners' English vowel mappings often involve one-to-multiple correspondences with both dominant and secondary mappings (Lee and Cho 2018, Tsukada et al. 2005), resisting straightforward categorization within the PAM-L2 typology and reflecting the less categorical nature of L2 vowel perception.

Among the most extensively studied contrasts are the English tense-lax pairs /i-ɪ/, /ɛ-æ/, and /u-ʊ/, which are frequently cited as problematic for Korean learners (Hong 2007, 2012, Lee and Baek 2025, Lee and Cho 2015). These contrasts tend to be perceptually assimilated to single Korean vowel categories: /i/ and /ɪ/ to Korean /i/, /ɛ/ and /æ/ to Korean /ɛ/, and /u/ and /ʊ/ to Korean /u/ (Hong 2007, Yang 1996), making it difficult for learners to perceive them as distinct phonemes. Beyond these frequently studied pairs, certain non-high vowel contrasts also pose considerable difficulty due to phonetic overlap with L1 categories, yet they have received comparatively less attention in the literature. In particular, /ɑ-ʌ/ and /ɔ-oʊ/ have been identified as problematic in the few studies that have examined them. Lee and Hwang (2016) reported that both /ɑ/ and /ʌ/ were frequently misidentified as /ɑ/, /ʊ/, or /ʌ/ even after HVPT training, and that /ɔ/ and /oʊ/ were similarly confused with /ɔ/, /oʊ/, and /ʌ/. Hong (2007) likewise noted that the /ɑ-ʌ-ɔ/ triplet is among the most perceptually confusing sets for Korean learners. Supporting this observation, Tsukada et al. (2005) found that Korean listeners' perceptual mapping of these vowels to L1 categories was highly variable: while /ɑ/ was predominantly mapped onto Korean /a/, /ʌ/ was dispersed across multiple Korean categories /a/, /ʌ/, /o/, and /ɨ/. These findings collectively suggest that /ɑ-ʌ/ and /ɔ-oʊ/ warrant greater research attention alongside the more commonly studied tense-lax pairs. In addition, /ɛ-eɪ/ was included to maintain continuity with previous HVPT research on Korean learners' English vowel training and to provide a relatively high-accuracy contrast that could function as a methodological baseline within the training set.

Previous studies have also shown that perceptual difficulty is not uniform across vowel contrasts. For example, Hong (2012) found that Korean learners' identification accuracy differed systematically across pairs, with /i-ɪ/ identified more accurately than /u-ʊ/, which in turn was more accurate than /ɛ-æ/, and that high vowels were generally better identified than non-high vowels. Lee and Hwang (2016) similarly reported different learnability patterns across multiple English vowel pairs. These findings indicate that not all English vowels present equal difficulty for Korean learners and that perceptual learning may proceed differently depending on the phonetic

relationship between L1 and L2 categories (Lee and Baek 2025). Given this variability, comparing AI-generated and human-recorded training input requires the inclusion of a broad range of vowel contrasts, allowing for a more complete evaluation of whether the two training conditions produce consistent effects across the full spectrum of difficulty levels.

Taken together, these studies indicate that L2 vowel perception is strongly shaped by learners' L1 phonological system (Ingram and Park 1997), and that language experience and exposure to L2 input can facilitate gradual improvement in perceptual accuracy over time (Flege et al. 1997). However, in EFL contexts where natural exposure to diverse L2 input is limited, such improvement is unlikely to occur through everyday experience alone. Structured perceptual training, such as HVPT, may therefore play an important role in providing the systematic and varied input necessary for developing accurate L2 vowel perception.

2.3 High Variability Phonetic Training (HVPT)

High Variability Phonetic Training (HVPT) provides a structured approach to promoting L2 phonetic learning by exposing learners to diverse speech input (Logan et al. 1991, Thomson 2012). HVPT typically involves presenting multiple speech tokens produced by different talkers so that learners encounter variability in acoustic realization. Recent meta-analytic evidence indicates that HVPT produces small-to-medium improvements in L2 speech perception among EFL learners (Choe et al. 2025), and another meta-analysis of 79 studies also reported medium-to-large effects, including evidence of long-term retention and generalization to novel stimuli (Uchihara et al. 2025). Generalization is widely considered as a central outcome of phonetic training, as it suggests that learners have developed more robust phonetic categories rather than memorizing specific training stimuli (Bradlow et al. 1999, Lively et al. 1993). Previous HVPT studies have also reported successful generalization across novel talkers and contexts (Carlet and Cebrian 2014, Leong et al. 2018, Zhang et al. 2021).

Although HVPT was originally designed as a perceptual training paradigm (Logan et al. 1991), its effects have been shown to extend to production (Bradlow et al. 1997). A meta-analysis by Sakai and Moorman (2018) reported that perceptual training produced medium-sized effects on perception and small-sized effects on production. However, the transfer from perception to production is not always robust. Hwang and Lee (2015), for instance, found minimal production gains from perceptual training alone and recommended combining perceptual and production training to maximize learning outcomes. Findings comparing perception-based and production-based training approaches have also been mixed. Li (2024) reported that a production-based training group outperformed a perception-based group in both perception and production, whereas Lee et al. (2020) found that perception-based instruction yielded larger gains than production-based instruction. Given these mixed findings, combining perceptual and production components within a single training program may represent a practical approach to addressing both modalities, particularly in classroom settings.

In line with this direction, recent scholarship has called for modifications to canonical HVPT to better suit classroom contexts. Thomson (2018) noted that despite its strong empirical support, HVPT has not been widely adopted in L2 classrooms, and encouraged more flexible, less laboratory-like implementations that can sustain learner engagement. Wong (2026) similarly advocated for adapting HVPT to meet diverse learner needs. Building on these recommendations, the present study adopted a classroom-based approach to HVPT in which the core framework of high-variability perceptual training was maintained while incorporating additional elements such as production practice through ASR-based speaking tasks, review game activities, and adaptive task adjustments based on learner performance. Although these modifications change how HVPT is delivered in the classroom, the core principle remains the same: learning is driven by variability in the input.

Variability is regarded as a key factor in phonetic training because exposure to multiple speakers and varied acoustic realizations helps learners attend to relevant phonetic cues while ignoring irrelevant variability in the signal (Brekelmans et al. 2025, Thomson 2012). However, variability is not limited to the number of speakers but also includes the acoustic characteristics of the input that shape how phonetic categories are learned. In this sense, examining how different types of speech input, such as AI-generated versus human-produced speech, affect phoneme identification, generalization across words and voices, and contrast-specific learning patterns is theoretically meaningful, as these input types may embody different forms of variability that influence phonetic category learning.

2.4 AI-Generated Speech in L2 Pronunciation and Perception Training

Recent studies have begun to explore the pedagogical use of AI-based tools in language learning. Research suggests that AI-supported pronunciation tools can improve pronunciation accuracy, increase learner confidence, and enhance engagement (Mohammadkarimi 2024), while meta-analytic evidence indicates that AI-based tools can support improvements in intelligibility and learner motivation (Vancova 2023). Among the AI-based tools, neural text-to-speech (TTS) systems have substantially improved the naturalness and intelligibility of synthetic speech (Bione and Cardoso 2020, Shen et al. 2018) and have increasingly been used as an effective pedagogical resource for L2 learning (Bione and Cardoso 2020, Liakin et al. 2017). In the present study, AI-generated speech refers to speech produced by neural TTS systems, which can generate highly natural and controllable synthetic voices. AI-generated speech offers several advantages for language instruction, including flexible voice control (e.g., adjustments in pitch, accent, and speaker identity), high accessibility through unlimited input availability (Bione and Cardoso 2020), and increasing levels of perceptual realism (Barrington et al. 2025).

Despite these advances, AI-generated and human speech differ in the source and nature of their acoustic variability, with direct implications for HVPT input. Modern neural TTS models the text-to-speech mapping as a one-to-many problem in which a single text corresponds to many possible acoustic realizations along pitch, duration, energy, and prosodic dimensions, and they handle this multimodality through speaker conditioning and variance predictors rather than through physiological articulation (Tan et al. 2021). When such variance information is limited, TTS tends to produce smoothed mel-spectrograms that approximate the average distribution rather than capturing token-to-token variation, a tendency known as the over-smoothness problem and confirmed by markedly lower acoustic variation indices in synthesized than in natural speech (Ren et al. 2022). Human speech, by contrast, embodies micro-variability arising from articulatory routines, coarticulatory effects, and speaker-specific anatomical differences, yielding broader within- and between-speaker acoustic distributions. Direct comparisons corroborate this distinction. John and Cardoso (2016) found that contemporary TTS output is segmentally indistinguishable from a native speaker for problematic English consonants and for the tense-lax vowel contrasts /i-ɪ/, /u-ʊ/, and /ɛ-æ/, yet has difficulty with non-default sentence stress because TTS lacks the semantic understanding that drives prosodic flexibility. Perceptual studies show a similar pattern: Smith et al. (2016) reported that human speech significantly outperformed TTS on comprehensibility, naturalness, accuracy, and intelligibility, while Bione and Cardoso (2020) found that the difference was mainly in naturalness, with intelligibility and comprehensibility statistically equivalent across the two voice types. Together, these findings indicate that AI-generated speech provides comparable segmental fidelity and functional comprehensibility but a more constrained variability profile, particularly at suprasegmental levels, than human speech. This distinction is consequential for HVPT, whose effectiveness rests on systematic variability that enables learners to extract robust phonetic category representations across speakers and contexts (Lively et al. 1993). Whether the more constrained

variability of AI voices nonetheless supplies functionally sufficient input for L2 phonetic learning remains an open empirical question that the present study addresses.

However, relatively few studies have examined the role of AI-generated speech as training input in phonetic training. Some HVPT studies have used acoustically manipulated or synthesized stimuli to facilitate phonetic learning (e.g., Cheng et al. 2019), but these studies primarily focused on artificially modified speech with cue manipulation rather than naturally generated AI speech, focusing on redirecting learners' cue weighting. By far only one recent study by Al-Shami and Cardoso (2025) incorporated naturally generated TTS voices in HVPT-inspired training, but it focused on learners' phonological awareness of English past tense allomorph rather than phoneme perception or production outcomes. Moreover, their study compared different types of synthetic speech rather than examining whether AI-generated speech can effectively replace human-recorded speech in HVPT. Consequently, the effectiveness of AI-generated speech as a primary source of training input in HVPT framework remains largely unexplored. As outlined above, AI-generated and human speech embody different forms of acoustic variability, which may influence not only the magnitude of perceptual learning but also patterns of generalization. The present study addresses this gap by directly comparing AI-generated and human-recorded speech within an HVPT-inspired pronunciation training for English vowel contrasts. Specifically, the study investigates three research questions:

1. Does pronunciation training with AI-generated speech result in significantly greater improvement in vowel phoneme identification accuracy than pronunciation training with human-recorded speech?
2. Do the generalization patterns across voice type (familiar vs. new) and word type (trained vs. untrained) differ between the two training modalities?
3. Does the effect of training on phoneme identification accuracy vary across vowel pairs, and does this variation differ by training conditions?

3. Method

3.1 Participants

Seventy-four Korean EFL middle school students participated in the study. All participants were 15-year-old ninth graders who volunteered to take part in the research. They self-reported normal hearing and no known history of speech or language disorders. Participants were randomly assigned by class to one of two training conditions: an AI-voice group (n = 34; 17 males, 17 females) and a human-voice group (n = 40; 18 males, 22 females). The two groups did not differ significantly in pre-test phoneme identification accuracy, measured using the two-alternative forced-choice task described in Section 3.4 (β = -0.15, SE = 0.188, z = -0.8, p = 0.425), indicating comparable baseline proficiency prior to training.

3.2 Target Vowel Contrasts and Stimuli

Six English vowel contrasts (/i-ɪ/, /ɛ-æ/, /ɛ-eɪ/, /ɑ-ʌ/, /ɔ-oʊ/, /u-ʊ/) were selected as training targets. The five vowel contrasts other than /ɛ-eɪ/ are commonly confused by Korean learners of English and served as the primary training targets of the present study. Although /ɛ-eɪ/ was expected to be relatively easy for learners, it was retained for methodological comparability with previous HVPT work and as a baseline contrast within the stimulus set.

Training stimuli were adapted from previous studies (Lee and Hwang 2016, Lee and Lyster 2017) with additional lexical items selected based on word frequency in the Corpus of Contemporary American English (Davies 2008-) and vocabulary lists from the participants' English textbooks (Lee et al. 2015) to ensure lexical familiarity. For each vowel contrast, eleven minimal pairs of mono- or disyllabic words were selected. As shown in Table 1, the minimal pairs numbered 1-10 served as training stimuli, whereas the minimal pair numbered 11 was reserved as an untrained test item and was not used during training. The /u-ʊ/ contrast yielded only eight training pairs because minimal pairs for this contrast are scarce in English. Two of these, *loo–look* and *boo–book*, are not strict minimal pairs but were included because they contrast the target vowels clearly and afford useful practice. Each training pair was also embedded in carrier sentences that differed only in the target word to provide more natural and contextualized pronunciation practice. Testing stimuli consisted only of isolated words without sentence contexts, including one trained minimal pair and one untrained minimal pair per contrast to examine generalization to novel lexical items. The full list of stimuli is provided in Table 1.

Table 1. List of Training and Testing Stimuli

	/i-ɪ/	/ɛ-æ/	/ɛ-eɪ/	/ɑ-ʌ/	/ɔ-oʊ/	/u-ʊ/
1	**steal-still**	**said-sad**	**men-main**	**cot-cut**	**call-coal**	**shoed-should**
2	beat-bit	set-sat	let-late	hot-hut	saw-sew	luke-look
3	teen-tin	bed-bad	tell-tail	stock-stuck	paul-pole	loo-look
4	cheek-chick	lend-land	letter-later	lock-luck	bought-boat	boo-book
5	peel-pill	mess-mass	wet-wait	rob-rub	caught-coat	
6	leave-live	kept-capped	pepper-paper	collar-color	gall-hole	cooed-could
7	feel-fill	kettle-cattle	test-taste	cops-cups	ball-bowl	fool-full
8	sheep-ship	men-man	selling-sailing	log-lug	bald-bold	boots-puts
9	cheeks-chicks	pen-pan	debt-date	wander-wonder	cost-coast	suit-soot
10	bean-bin	letter-latter	bread-braid	pot-putt	pause-pose	
11	wheel-will	head-had	get-gate	not-nut	law-low	pool-pull

Note. Minimal pairs numbered 1-10 were used during training. **Boldface** indicates the trained test item selected from the training set, whereas underlining indicates the untrained test item, which was not used during training and appeared only in the pre- and post-test.

Based on these stimuli, two sets of audio materials were developed for the Human-voice and AI-voice groups. For the Human-voice group, four standard American English speakers (two male, two female) recorded the target stimuli in a quiet environment using a MacBook Pro built-in microphone (44.1 kHz). The recordings were noise-reduced by Audacity (Audacity Team 2025) and intensity-normalized to 70dB in Praat (Boersma and Weenink 2022). For the AI-voice group, stimuli were generated using an AI text-to-speech platform (LOVO 2023) with four synthetic American English voices (two male, two female). The AI voices were selected by the first researcher through auditory judgment, based on their perceived naturalness and general comparability to the human speakers in gender and overall voice quality. The selection was intended to keep the two training conditions as comparable as possible while differing primarily in the source of the auditory input, namely AI-generated versus human-recorded speech. The AI voices were thus chosen perceptually rather than matched to the human speakers on measured acoustic parameters (e.g., F0, speaking rate, or vowel duration). The first author segmented all stimuli and integrated them into the training materials.

3.3 Test Materials and Generalization Conditions

Test stimuli were made using the same procedures as the training materials for both human and AI voices. Pre- and post-tests measured vowel identification accuracy using a two-alternative forced-choice (2AFC) phoneme identification task. In each trial, participants heard a stimulus word three times and selected the corresponding option (e.g., *steal* and *still)*. There were 48 items in each test, designed to examine generalization across voice type and lexical items (6 vowel contrasts × 2 minimal pair words × 2 voices (familiar / new) × 2 word types (trained / untrained)). To assess voice generalization, half of the stimuli were produced by human voices and the other half by AI voices. For the AI-voice group, AI voices were the familiar voice type and human voices were new voice type; the reverse applied to the Human-voice group. Word-type generalization was examined by trained words used during training and untrained words responding novel lexical items with the same vowel contrasts. These conditions allowed examination of whether perceptual improvement generalized beyond specific talkers and lexical items. Students' responses were coded as binary outcomes, with correct responses coded as 1 and incorrect responses as 0.

3.4 Procedure

The experiment followed a pre-test–training–post-test design. Students who volunteered to participate in the HVPT-informed pronunciation training were assigned to either the AI-voice or Human-voice group. Prior to training, participants completed a pre-test measuring their English vowel identification accuracy. Training was conducted over eight weeks with two sessions per week (16 sessions in total), each lasting approximately 25 minutes. Each session focused on one vowel contrast. Training was implemented through structured classroom activities, as summarized in Table 2. Each session included explicit instruction with articulatory diagrams, listen-and-repeat with group-specific stimuli, and 2AFC identification tasks delivered via Moodle LMS with immediate feedback (Moodle 2025). Learners also practiced producing target words and sentences using ASR-enabled mobile applications. Supplementary worksheets and game activities (Kahoot! 2025) were used in review sessions to reinforce learning. Instructional procedures and feedback were identical across groups, differing only in the source of the audio stimuli.

Table 2. Training Materials

Activity	Materials	Stimuli
Teacher's explicit explanation	Articulatory diagrams	words
Listen and Repeat	Stimuli audio files	words, sentences
Identification practice	2AFC identification tasks via Moodle LMS	words
Production practice	ASR-enabled apps	words, sentences
Review	Online Kahoot games	words, sentences

Sessions began with the teacher's explicit instruction using articulatory diagrams. The teacher used diagrams to show how the two vowels in each vowel contrast differed articulatorily and briefly explained the positions of the articulators, including mouth shape and tongue position. This was followed by listen-and-repeat practice using group-specific stimuli (AI or human voices). The teacher played the AI-generated or human-recorded stimuli for the minimal pairs containing the target vowel contrast, and students listened to and repeated the words. Students also worked in pairs, taking turns reading words or carrier sentences and identifying which word or sentence their partner had produced.

Participants then practiced vowel identification individually through 2AFC tasks delivered via Moodle. Each session included 40 identification trials (10 training words × 4 talkers), in which they listened to the group-specific training stimulus and selected the correct option from two minimal-pair options. Immediate feedback was provided after each response, allowing learners to confirm whether their choice was correct. Because the study was implemented in a classroom setting, a criterion-based practice procedure was adopted. Students were encouraged to reach at least 60% accuracy during each training session; if this threshold was not met, they were encouraged to complete additional identification practice. This procedure was applied equally across the AI-voice and Human-voice groups and was intended to ensure a minimal level of engagement with the target contrast, while the primary between-group manipulation remained the source of the auditory input, namely AI-generated versus human-recorded speech. Specifically, this criterion-supported procedure was designed to prevent the training effects from being underestimated due to low task engagement, insufficient participation, or inattentive responding by some learners. It also served as a pedagogically necessary support in the classroom context, where encouraging active learner participation was an important part of implementing the training. In practice, however, additional practice was rarely needed; therefore, this procedure was unlikely to have affected the overall amount of training exposure across learners.

After the individual identification practice with Moodle, learners proceeded to practice producing the target vowels using automatic speech recognition (ASR) tools. This production component was included to provide articulatory practice with the same target vowel contrasts while making the training more suitable for classroom-based pronunciation instruction. Using the speech-to-text (STT) function in Google Docs or Google Translate on their individual educational tablet devices, students pronounced the given words and sentences and compared their intended pronunciation with the machine-generated transcription.

In addition, engagement-enhancing activities were incorporate to reduce the monotony of repeated HVPT identification practice and to make the training more suitable for middle school learners. After every three training sessions, a review session was conducted using Kahoot activities to reinforce previously learned contrasts. These activities followed the same basic 2AFC identification format as the Moodle tasks, but were implemented as real-time online games in which students competed together and accumulated points. Students selected the word or carrier sentence that matched the audio stimulus played by the teacher and received points for correct responses.

Across the training period, each vowel contrast was taught twice using different lexical items (see Table 3). In the first round (Sessions 1-8), minimal pairs 1-5 in Table 1 were taught, whereas in the second round (Sessions 9-16), minimal pairs 6-10 in Table 1 were taught, resulting in a total of 10 minimal pairs per contrast. Both groups completed the same number of sessions, used identical lexical items, and followed the same instructional procedures, differing only in the source of the speech stimuli. Following the completion of training, participants completed the post-test, which was structurally identical to the pre-test.

Table 3. Training Plan

Session	1	2	3	4	5	6	7	8
Vowel contrast	/i-ɪ/	/ɛ-æ/	/ɛ-eɪ/	review	/ɑ-ʌ/	/ɔ-oʊ/	/u-ʊ/	review
Session	9	10	11	12	13	14	15	16
Vowel contrast	/i-ɪ/	/ɛ-æ/	/ɛ-eɪ/	review	/ɑ-ʌ/	/ɔ-oʊ/	/u-ʊ/	review

3.5 Data Coding and Statistical Analysis

All responses in pre- and post-test were coded at the item level as binary outcomes (correct = 1, incorrect = 0). Data were analyzed using binomial generalized linear mixed-effects models (glmer) (Baayen et al. 2008)

implemented in R (R Core Team 2025) using the lme4 package (Bates et al. 2015). Fixed effects included Group (AI vs. Human) and Time (pre vs. post), along with Voice Type, Word Type, and Vowel Pair depending on the research question, and their interactions. Random intercepts for participants and items were included, and random slopes for participants were specified. The significance of fixed effects was evaluated using Type III Wald chi-square tests implemented in the car package (Fox and Weisberg 2019). Predicted probabilities and pairwise contrasts were estimated using the emmeans package (Lenth and Piaskowski 2026) to examine patterns of improvement and generalization.

4. Results

Descriptive statistics for overall phoneme identification accuracy by group and time are presented in Table 4. Accuracy was calculated as the participant-level mean percentage of correct responses. These descriptive statistics provide an overview of the pre-post changes before the GLMM results are reported. Additional descriptive statistics for the more specific analysis conditions, including Group × Time × Voice Type, Group × Time × Word Type, and Group × Time × Vowel Pair, are provided in Appendix A.

Table 4. Descriptive Statistics for Overall Phoneme Identification Accuracy by Group and Time

Group	Time	*M* (% accuracy)	*SD*	*n*
Human	Pre	74.9	10.4	40
	Post	82.6	8.7	40
AI	Pre	72.5	12.7	34
	Post	84.4	7.6	34

Note. *M* and *SD* are reported in percentage points based on participant-level mean accuracy scores.

4.1 Overall Training Effects on Phoneme Identification (RQ1)

The binomial generalized linear mixed model was conducted to examine the effects of Group (AI vs. Human), Time (pre vs. post), and their interaction on phoneme identification accuracy (Table 5). The main effect of Group was not significant (β = -0.150, *SE* = 0.188, *z* = -0.80, *p* = .425), indicating no baseline difference between the two groups at pre-test. A significant main effect of Time was observed (β = 0.543, *SE* = 0.105, *z* = 5.15, *p* < .001), suggesting overall improvement from pre- to post-test. Importantly, the Group × Time interaction reached statistical significance (β = 0.303, *SE* = 0.153, *z* = 1.98, *p* = .048), indicating that the AI-voice group demonstrated greater pre-post gains compared to the Human-voice group. To further clarify this interaction, planned contrasts based on the estimated marginal means (Table 6) revealed significant pre-post improvement in both groups, with a larger gain observed in the AI-voice group. The Human-voice group increased from 83.0% at pre-test to 89.3% at post-test (Δ = 6.3 percentage points), whereas the AI-voice group improved from 80.8% to 90.7% (Δ = 9.9 percentage points), reflecting an additional 3.6 percentage-point gain in the AI-voice condition. The interaction pattern is illustrated in Figure 1.

Table 5. Binomial GLMM Results for Group × Time Effects on Phoneme Identification Accuracy

Fixed effect	β	*SE*	*z*	*p*
(Intercept)	1.584	0.241	6.59	< .001***
Group(AI)	-0.150	0.188	-0.80	.425
Time(Post)	0.543	0.105	5.15	< .001***
Group(AI):Time(Post)	0.303	0.153	1.98	.048*

Note. The Human-voice group and pre-test were used as reference levels. Coefficients are reported on the log-odds scale.
* *p* < .05, ** *p* < .01, *** *p* < .001.

Table 6. Predicted Probabilities and Percentage-point Changes from Pre- to Post-test

Group	Pre	Post	*Δ (pp)*	*p*
Human	0.830	0.893	6.3	< .001***
AI	0.808	0.907	9.9	< .001***

Note. Predicted probabilities were obtained from the fitted GLMM using estimated marginal means. *Δ (pp)* represents percentage-point changes from pre to posttest. *p*-values are from model-based pre-post contrasts (tests performed on the logit scale). * $p < .05$, ** $p < .01$, *** $p < .001$.

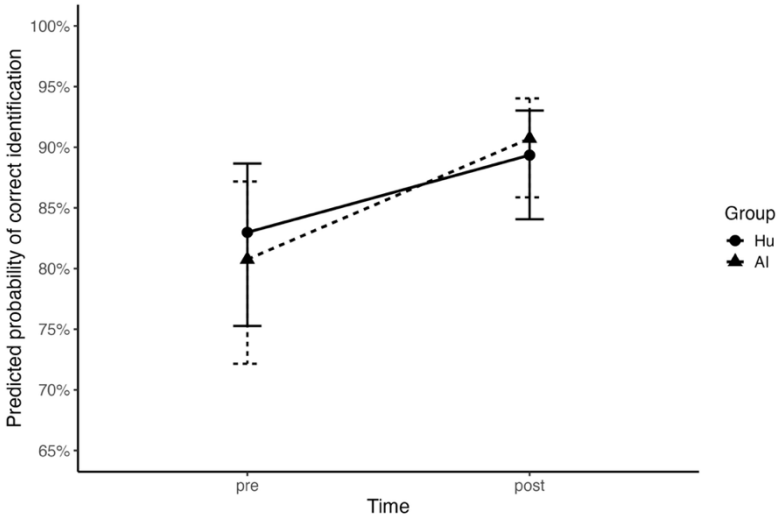

Figure 1. Predicted Probabilities of Correct Vowel Identification at Pre-test and Post-test for the AI-voice and Human-voice Groups

Note. Error bars indicate 95% confidence intervals.

4.2 Generalization Effects across Voice and Word (RQ2)

4.2.1 Generalization for voice type (new vs. familiar)

A binomial generalized linear mixed-effects model was conducted to examine the effects of Group, Time, and Voice Type (familiar vs. new) on phoneme identification accuracy (see Table 7). The main effect of Time was significant ($\beta = 0.430$, $SE = 0.136$, $z = 3.18$, $p = .001$), indicating overall improvement from pre- to post-test. The Group × Time interaction was also significant ($\beta = 0.448$, $SE = 0.201$, $z = 2.23$, $p = .026$), suggesting that the AI-voice group demonstrated overall greater pre-post gains than the Human-voice group. Planned contrasts on the estimated marginal means (Table 8) indicated both groups demonstrated significant improvement in the new voice condition, indicating successful generalization. AI-voice group showed numerically larger improvement ($\Delta = 10.3$ percentage points) than the Human-voice group ($\Delta = 5.0$ percentage points). However, the non-significant three-way interaction among Group, Time, and Voice Type ($\beta = -0.293$, $SE = 0.261$, $z = -1.12$, $p = .262$) suggests that the magnitude of pre-post improvement did not differ as a function of voice type across the two training groups. Figure 2 illustrates pre-post changes in predicted probabilities for new and familiar voice conditions across the two training groups.

Table 7. Binomial GLMM Results for Group × Time × VoiceType Effects on Phoneme Identification Accuracy

Fixed effect	*β*	*SE*	*z*	*p*
(Intercept)	1.642	0.321	5.12	< .001***
Group(AI)	-0.220	0.454	-0.48	.629
Time(Post)	0.430	0.136	3.18	.001**
Voice(Familiar)	-0.115	0.421	-0.27	.784
Group(AI)*Time(Post)	0.448	0.201	2.23	.026*
Group(AI)*Voice(Familiar)	0.139	0.826	0.17	.866
Time(Post)*Voice(Familiar)	0.230	0.175	1.31	.189
Group(AI)*Time(Post)*Voice(Familiar)	-0.293	0.261	-1.12	.262

Note. The Human-voice group, pre-test, and new voice were specified as reference levels. Coefficients are reported on the logit scale. * $p < .05$, ** $p < .01$, *** $p < .001$.

Table 8. Predicted Probabilities and Percentage-point Changes from Pre- to Post-test by Group and Voice Type

Group	Pre	Post	*Δ (pp)*	*p*
New Voice type				
AI	0.806	0.909	10.3	< .001***
Human	0.838	0.888	5.0	.002**
Familiar Voice type				
AI	0.809	0.906	9.6	< .001***
Human	0.822	0.899	7.8	< .001***

Note. Predicted probabilities were obtained from the fitted GLMM using estimated marginal means. *Δ (pp)* represents percentage-point changes from pre- to post-test. *p*-values are from model-based pre-post contrasts (tests performed on the logit scale). * $p < .05$, ** $p < .01$, *** $p < .001$.

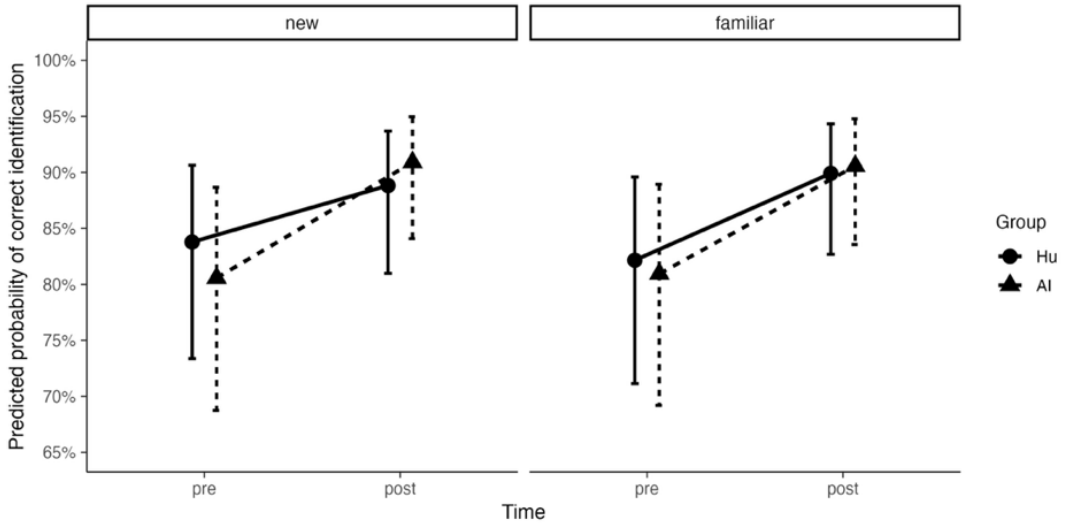

Figure 2. Predicted Probabilities of Correct Vowel Identification at Pre-test and Post-test by Group (AI-voice vs. Human-voice) across Voice Conditions (New vs. Familiar)

Note. Error bars represent 95% confidence intervals.

4.2.2 Generalization for word type (trained vs. untrained)

To examine generalization across lexical item types (trained vs. untrained words), a binomial GLMM was conducted with Group, Time, and WordType as fixed effects (Table 9). A significant main effect of Time was observed (β = 0.365, *SE* = 0.136, *z* = 2.68, *p* = .007), confirming overall improvement from pre- to post-test. Moreover, the Time × WordType interaction was significant (β = 0.362, *SE* = 0.175, *z* = 2.06, *p* = .039), suggesting that the improvement was greater for trained than untrained items. To further examine the generalization to the untrained lexical items, predicted probabilities were extracted from the model for each word-type condition (Table 10). Both groups demonstrated significant pre-post improvement in the untrained-word condition, indicating successful lexical generalization. In addition, the Group × Time interaction was significant (β = 0.417, *SE* = 0.202, *z* = 2.06, *p* = .039), indicating greater overall improvement in the AI-voice group relative to the Human-voice group. However, because the three-way interaction was not significant (β = -0.238, *SE* = 0.260, *z* = -0.91, *p* = .361), this pattern was comparable across the two training groups. Figure 3 illustrates the pre-post changes in predicted probabilities for trained and untrained word conditions across the two training groups.

Table 9. Binomial GLMM Results for Group × Time × Word type Effects on Phoneme Identification Accuracy

Fixed effect	β	*SE*	*z*	*p*
(Intercept)	1.654	0.320	5.17	< .001***
Group(AI)	-0.102	0.208	-0.49	.624
Time(Post)	0.365	0.136	2.68	.007**
WordType(Trained)	-0.137	0.421	-0.33	.744
Group(AI)*Time(Post)	0.417	0.202	2.06	.039*
Group(AI)*WordType(Trained)	-0.092	0.171	-0.54	.589
Time(Post)*WordType(Trained)	0.362	0.175	2.06	.039*
Group(AI)*Time(Post)*WordType(Trained)	-0.238	0.260	-0.91	.361

Note. The Human-voice group, pre-test, and untrained word were specified as reference levels. Coefficients are reported on the logit scale. * $p < .05$, ** $p < .01$, *** $p < .001$.

Table 10. Predicted Probabilities and Percentage-point Changes from Pre- to Post-test by Group and Word Type

Group	Pre	Post	Δ *(pp)*	*p*
Untrained WordType				
AI	0.825	0.912	8.6	< .001***
Human	0.839	0.883	4.3	.007**
Trained WordType				
AI	0.790	0.903	11.3	< .001***
Human	0.820	0.904	8.4	< .001***

Note. Predicted probabilities were obtained from the fitted GLMM using estimated marginal means. Δ *(pp)* represents percentage-point changes from pre- to post-test. *p*-values are from model-based pre-post contrasts (tests performed on the logit scale). * $p < .05$, ** $p < .01$, *** $p < .001$.

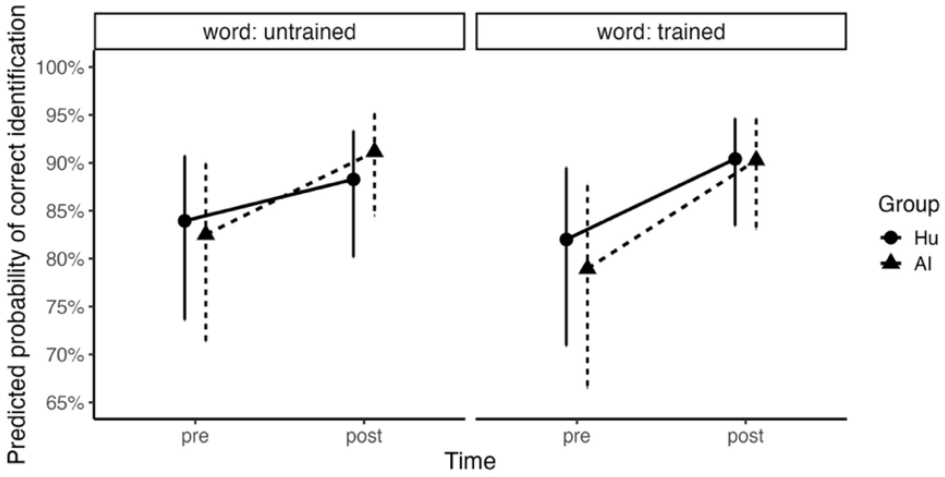

Figure 3. Predicted Probabilities of Correct Vowel Identification at Pre-test and Post-test by Group (AI-voice vs. Human-voice) across Word Types (Untrained vs. Trained)

Note. Error bars represent 95% confidence intervals.

4.3 Vowel Pair-Specific Training Effects (RQ3)

To address RQ3, a Type III Wald chi-square test was additionally conducted to evaluate the overall contribution of each fixed effect, especially with those included Pair with 6 levels. In Table 11, the main effect of Pair was significant, $\chi^2(5) = 70.77$, $p < .001$, suggesting baseline differences in identification accuracy across vowel pairs. The Time × Pair interaction was also significant, $\chi^2(5) = 75.11$, $p < .001$, indicating that the magnitude of pre-post improvement differed across vowel pairs. However, the three-way interaction among Group, Time, and Pair was not significant, $\chi^2(5) = 4.63$, $p = .463$, suggesting that the pattern of pair-specific improvement did not significantly differ between the AI and Human training conditions. Table 12 presents the fixed-effects estimates including each pair-specific fixed-effects. Consistent with the omnibus tests, several Time × Pair coefficients were significant, indicating that pre-post changes for specific vowel pairs deviated significantly from the average improvement across pairs. Specifically, significant deviations were observed for /ɑ-ʌ/ ($\beta = -1.042$, $p < .001$), /ɛ-æ/ ($\beta = 1.016$, $p < .001$), /i-ɪ/ ($\beta = -0.433$, $p = .045$), and /ɔ-oʊ/ ($\beta = 0.558$, $p = .023$). These results indicate that the extent of improvement varied across vowel contrasts. Crucially, however, none of the Group × Time × Pair coefficients were significant (all ps > .18), confirming that pair-specific improvement patterns did not significantly differ between the AI and Human conditions.

Table 11. Type III Wald Chi-square Test

	Chisq	*df*	*Pr(>Chisq)*
(Intercept)	90.59	1	< .001***
Group	0.93	1	.334
Time	11.34	1	< .001***
Pair	70.77	5	< .001***
Group:Time	3.44	1	.063
Group:Pair	5.26	5	.386
Time:Pair	75.11	5	< .001***
Group:Time:Pair	4.63	5	.463

Note. * $p < .05$, ** $p < .01$, *** $p < .001$.

Table 12. Fixed Effect Estimates from the GLMM

Fixed effect	*β*	*SE*	*z*	*p*
(Intercept)	1.681	0.177	9.52	< .001***
Group(AI)	-0.202	0.210	-0.97	.334
Time(Post)	0.488	0.145	3.37	< .001***
Pair(/ɑ-ʌ/)	-0.569	0.256	-2.22	.026*
Pair(/ɛ-æ/)	-1.604	0.252	-6.37	< .001***
Pair(/ɛ-eɪ/)	2.959	0.440	6.72	< .001***
Pair(/i-ɪ)/	0.016	0.263	0.06	.951
Pair(/ɔ-oʊ/)	0.018	0.263	0.07	.946
Group(AI)*Time(Post)	0.392	0.211	1.86	.064
Group(AI)*Pair(/ɑ-ʌ/)	-0.302	0.198	-1.53	.127
Group(AI)*Pair(/ɛ-æ/)	0.139	0.189	0.74	.461
Group(AI)*Pair(/ɛ-eɪ/)	-0.292	0.491	-0.60	.552
Group(AI)*Pair(/i-ɪ)/	0.098	0.215	0.46	.649
Group(AI)*Pair(/ɔ-oʊ/)	0.107	0.215	0.50	.620
Time(Post)*Pair(/ɑ-ʌ/)	-1.042	0.198	-5.27	< .001***
Time(Post)*Pair(/ɛ-æ/)	1.016	0.206	5.04	< .001***
Time(Post)*Pair(/ɛ-eɪ/)	-0.507	0.526	-0.97	.335
Time(Post)*Pair(/i-ɪ)/	-0.433	0.216	-2.00	.045*
Time(Post)*Pair(/ɔ-oʊ/)	0.558	0.246	2.27	.023*
Group(AI)*Time(Post)*Pair(/ɑ-ʌ/)	0.169	0.290	0.58	.560
Group(AI)*Time(Post)*Pair(/ɛ-æ/)	0.256	0.310	0.83	.408
Group(AI)*Time(Post)*Pair(/ɛ-eɪ/)	0.462	0.778	0.59	.553
Group(AI)*Time(Post)*Pair(/i-ɪ)/	-0.422	0.317	-1.33	.183
Group(AI)*Time(Post)*Pair(/ɔ-oʊ/)	-0.440	0.357	-1.23	.218

Note. * $p < .05$, ** $p < .01$, *** $p < .001$. Pair was sum-coded. The remaining pair (/u-ʊ/) is represented by the negative sum of these coefficients. Group (Human) and Time (Pre) were treatment-coded as reference levels. Coefficients are reported in log-odds.

Given the absence of a significant three-way interaction, we next examined the magnitude of improvement within each training condition to address the second part of RQ3, namely, which vowel pairs showed greater gains under each condition. Table 13 presents the predicted probabilities at pre- and post-test, the corresponding percentage-point changes (*Δ*), and 95% confidence intervals for each vowel pair, ranked by the size of improvement within each group. In the Human group, the largest improvement was observed for /ɛ-æ/ (*Δ* = 31.01 percentage points, 95% CI [21.7, 40.32]), followed by /u-ʊ/ (*Δ* = 15.01, 95% CI [7.02, 23]) and /ɔ-oʊ/ (*Δ* = 9.42, 95% CI [3.44, 15.4]). Minimal or no improvement was found for /i-ɪ/ and /ɛ-eɪ/, and performance on /ɑ-ʌ/ showed a decrease from pre- to post-test (*Δ* = -11.66, 95% CI [-19.76, -3.56]). A highly similar pattern emerged in the AI group. The greatest gain was again observed for /ɛ-æ/ (*Δ* = 39.37, 95% CI [28.23, 50.51]), followed by /u-ʊ/ (*Δ* = 18.51, 95% CI [9.5, 27.52]) and /ɔ-oʊ/ (*Δ* = 9.84, 95% CI [3.25, 16.44]). The other vowel pairs showed minimal change, and no substantial improvement was observed for /ɑ-ʌ/. Overall, both training conditions yielded comparable improvement patterns across vowel pairs, with /ɛ-æ/ consistently showing the largest gains and /ɑ-ʌ/ showing little or negative change. These results suggest that while the training effects were vowel-specific, the relative hierarchy of improvement was largely similar across AI-voiced and human-voiced training. Figure 4 visually illustrates this pattern, showing vowel-specific variation in improvement without group-level divergence.

Table 13. Ranks of Vowel Pairs by Groups

Pair	Pre	Post	*Δ (pp)*	*CI* (95%)	Rank
Human group					
/ɛ-æ/	0.519	0.829	31.01	[21.7, 40.32]	1
/u-ʊ/	0.703	0.853	15.01	[7.02, 23]	2
/ɔ-oʊ/	0.845	0.940	9.42	[3.44, 15.4]	3
/i-ɪ/	0.845	0.852	0.70	[-4.89, 6.28]	4
/ɛ-eɪ/	0.990	0.990	-0.02	[-1.2, 1.16]	5
/ɑ-ʌ/	0.752	0.636	-11.66	[-19.76, -3.56]	6
AI group					
/ɛ-æ/	0.503	0.897	39.37	[28.23, 50.51]	1
/u-ʊ/	0.712	0.897	18.51	[9.5, 27.52]	2
/ɔ-oʊ/	0.832	0.931	9.84	[3.25, 16.44]	3
/ɛ-eɪ/	0.984	0.993	0.88	[-0.58, 2.34]	4
/i-ɪ/	0.831	0.834	0.33	[-6.07, 6.73]	5
/ɑ-ʌ/	0.647	0.649	0.15	[-8.9, 9.2]	6

Note. Predicted probabilities were obtained from the fitted GLMM using estimated marginal means. *Δ (pp)* represents percentage-point changes from pre- to post-test. *p*-values are from model-based pre-post contrasts (tests performed on the logit scale).

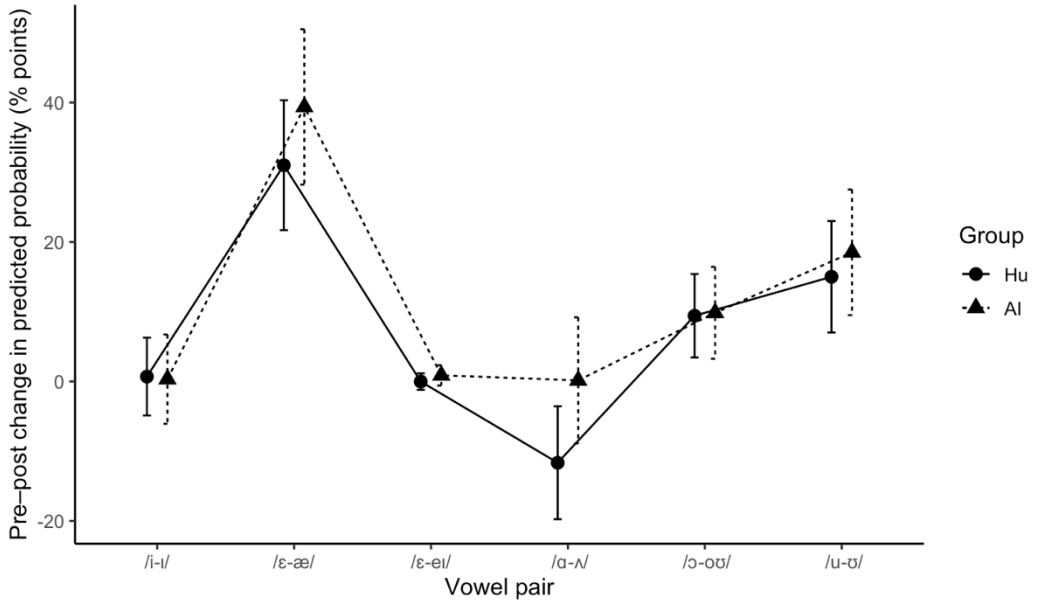

Figure 4. Pre-post Changes in Predicted Probabilities (% points) across Vowel Pairs by Training Group

Note. Error bars represent 95% confidence intervals.

5. Discussion

5.1 AI- vs. Human-voice: Learning Strength and Learning Structure

The most important finding of the present study is that AI-generated speech enhanced the magnitude of phonetic learning without altering the underlying structure of phonological acquisition. While the AI-voice group demonstrated significantly greater overall gains in phoneme identification accuracy compared to the Human-voice group, this advantage did not translate into qualitative differences in how learning generalized or how phonological difficulty was organized.

Across both voice and word-type generalization conditions, the absence of significant three-way interactions indicates that the pattern of generalization was comparable between training conditions. That is, although the AI group improved more in absolute terms, the way in which learning extended to new voices and untrained lexical items remained structurally similar to that observed in the Human-voice group. Importantly, vowel-pair-specific patterns further reinforce this interpretation: while certain contrasts showed larger gains than others, the relative hierarchy of difficulty across vowel pairs was preserved across groups. In other words, AI training did not reorganize the phonological difficulty landscape shaped by Korean L1 phonology; it amplified learning within that structure.

One factor that may contribute to this magnitude advantage concerns the acoustic properties of the input. The acoustic variability of AI-generated speech is shaped by model design rather than physiological articulation (Ren et al. 2022, Tan et al. 2021, see Section 2.4), and it is tentatively hypothesized that its potentially more controlled within-category distributions may facilitate more efficient extraction of phonetic category prototypes during perceptual learning. These acoustic differences between the two voice types, as reported in the previous literature, provide a plausible interpretation for the greater gains observed in the AI condition. Notably, the learning patterns found in the present study align with the direction predicted by this literature, which lends coherence to a variability-based account. It should be emphasized, however, that this account remains speculative: the variability profiles of the two stimulus sets were not acoustically verified in the present study, and the AI advantage could also reflect other differences between the stimulus sets. Direct acoustic analysis is therefore needed to substantiate this interpretation (see Limitations).

Taken together, these findings suggest a distinction between learning strength and learning structure. AI-generated input appears capable of strengthening category-level perceptual restructuring, resulting in larger performance gains, yet the mechanisms governing generalization and category-level organization remain consistent with those elicited by human speech input. From an SLM perspective, this implies that AI speech can effectively support perceptual restructuring processes (Bohn and Munro 2007) without modifying the fundamental architecture of L2 phonological learning. It suggests that AI-generated stimuli can provide variability sufficient for category restructuring, and that the perceptual system responds to such input in ways comparable to natural human speech. Thus, the contribution of AI-generated speech lies not in altering how learners generalize or reorganize phonetic categories, but in increasing the degree to which those processes succeed.

5.2 Generalization and the Variability of AI Input

The generalization findings provide further support for the interpretation that both AI- and human-based pronunciation training promoted category-level phonetic restructuring rather than item-specific learning. Improvement extended beyond the specific training stimuli, as learners demonstrated gains in both new voice type

conditions and untrained lexical items. Such transfer is widely regarded as a hallmark of category-level rather than item-specific phonetic learning within the Speech Learning Model (Bohn and Munro 2007), as it indicates that learners are not merely memorizing stimulus-specific acoustic patterns but are reorganizing their underlying phonetic representations (Bradlow et al. 1999, Lively et al. 1993). This pattern is consistent with previous meta-analytic findings indicating that HVPT can promote generalization to novel stimuli, particularly when training involves real-word stimuli and multiple talkers (Choe et al. 2025). Crucially, the absence of significant Group × Time × Voice Type and Group × Time × Word Type interactions suggests that the patterns of generalization, and thus the underlying learning processes, were similar across training modalities. In both groups, learners were able to abstract away from talker-specific acoustic variability and lexical familiarity, applying newly strengthened phonetic categories to novel input. This pattern aligns with the SLM prediction that sufficient phonetic variability facilitates the formation or restructuring of L2 categories by encouraging learners to attend to relevant acoustic cues rather than relying on L1-based assimilation strategies (Brekelmans et al. 2025, Thomson 2012).

These findings also bear on theoretical debates concerning the role of variability in perceptual learning. Traditional HVPT emphasized exposure to multiple speakers as a means of destabilizing L1-driven perceptual mappings and promoting more robust category formation (Lively et al. 1993). The present results suggest that AI-generated stimuli can provide variability sufficient to trigger these restructuring processes, even if the acoustic micro-variability may differ from that of natural human speech, as suggested in prior work (Kim and Seong 2025). That AI-trained learners generalized to new voices and untrained items indicates that the variability of the AI input was functionally sufficient to support the perceptual abstraction that HVPT is designed to elicit, although this functional sufficiency was inferred from learning outcomes rather than verified acoustically. In this sense, what appears to matter is not the human origin of the signal per se, but whether the input contains enough systematic contrastive information to drive perceptual recalibration. Taken together, the present findings suggest that the perceptual system treats AI-generated speech as functionally comparable to human-generated speech for the purposes of category-level perceptual restructuring.

5.3 Preservation of Vowel-Pair Difficulty Patterns Across Training Conditions

Although the relative pattern of vowel improvement did not differ between AI- and Human-voice groups, the magnitude of gains varied across vowel pairs. Across both groups, some contrasts showed substantial improvement (e.g., /ɛ-æ/, /u-ʊ/, and /ɔ-oʊ/), whereas others showed minimal change (e.g., /ɑ-ʌ/) or near-ceiling performance (e.g., /i-ɪ/).

Several of these contrast-specific patterns are consistent with findings reported in previous studies. As in Lee and Hwang (2016), learners showed near-ceiling performance for /ɛ-eɪ/ at pretest (accuracy approaching 99%), an anticipated result given that this contrast was included as a methodological baseline. Among the three most frequently studied contrasts, pretest accuracy followed the same easiness hierarchy reported by Kim (2010) and Hong (2012): /i-ɪ/ > /u-ʊ/ > /ɛ-æ/ in both groups. Training gains, however, followed the reverse order, with /ɛ-æ/ showing the largest improvement ($\Delta = 31.01$ in the Human group; $\Delta = 39.37$ in the AI group), followed by /u-ʊ/ ($\Delta = 15.01$; $\Delta = 18.51$). This pattern may reflect the status of /æ/ as a novel vowel for Korean learners, making it particularly responsive to perceptual training. As a result, posttest accuracy for these three contrasts converged to broadly comparable levels. The contrast /i-ɪ/ showed minimal improvement, which can be attributed to a near-ceiling effect at pretest.

Interestingly, /ɔ-oʊ/ showed pretest accuracy similar to that of /i-ɪ/ yet demonstrated considerable training gains. This may be because /ɔ-oʊ/ benefits from salient durational cues, as Korean learners often rely on temporal

information when distinguishing English vowel contrasts (Lee and Baek 2025). Just as the durational difference between /ɛ/ and /eɪ/ makes that pair perceptually transparent, training may have helped learners recognize a similar durational distinction in /ɔ-oʊ/. This interpretation is supported by the observation that posttest accuracy for /ɔ-oʊ/ reached levels comparable to /ɛ-eɪ/ (94.3% in the Human group; 93.5% in the AI group).

The contrast /ɑ-ʌ/ proved most resistant to training, consistent with previous literature identifying this pair as particularly difficult for Korean learners (Hong 2007, Tsukada et al. 2005). The AI group showed minimal change, while the Human group showed a slight decrease in accuracy, a pattern also reported in Lee and Hwang (2016). It should be noted, however, that the three-way interaction among Group, Time, and Vowel Contrast was not statistically significant, so the apparent difference between the two groups on this contrast should be interpreted with caution. With this caveat in mind, one tentative possibility is that the more variable nature of human speech may have introduced additional perceptual uncertainty for a contrast lacking stable L1 category mappings, whereas the more consistent AI input may have provided a more stable, though still insufficient, learning signal. As this interpretation was not verified through direct acoustic analysis, it remains speculative and awaits confirmation in future work.

Importantly, despite these contrast-specific differences, the relative hierarchy of difficulty across vowel pairs remained stable across AI- and human-based training. Although the AI-voice group showed numerically larger gains across all contrasts, the ranking of vowel-pair difficulty was preserved in both groups. This stability suggests that AI-generated speech does not fundamentally change the perceptual learning patterns or learners' response to L1-L2 phonological relationships. Instead, it helps learners improve specific phoneme distinctions within that system. In this respect, the results reinforce the interpretation advanced earlier: AI-voice training enhances the efficiency of phonetic learning without altering the underlying organization of L2 vowel perception**.**

5.4 Pedagogical Implications: The Potential and Boundaries of AI-Voice Training

The findings of the present study carry important pedagogical implications for the integration of AI-generated speech into L2 phonetic training. Most notably, the results suggest that classroom-based HVPT-informed vowel training using AI-generated input can yield learning gains that are comparable to, and in some cases greater than, those achieved through human-recorded input. This indicates that AI-generated speech is not merely a technological substitute but a pedagogically viable medium for promoting phonetic category development. Importantly, the preservation of generalization patterns and vowel improvement hierarchies across training modalities suggests that AI input engages the same perceptual learning mechanisms as human speech. This may tentatively imply that AI-generated materials can provide sufficient phonetic variability and contrastive information to promote the restructuring processes predicted by SLM. At the same time, the results do not suggest that AI fundamentally alters the nature of phonological learning. Rather than replacing human input in a qualitative sense, AI-generated input appears to preserve the way human speech input shapes phonetic learning while enhancing the efficiency of an otherwise shared learning mechanism.

Collectively, these findings suggest that AI-voice training may serve as a scalable, accessible supplement to traditional pronunciation training, particularly in contexts where access to multiple native speakers is limited. Neural TTS platforms enable training materials to be generated rapidly, at scale, at low cost, and with full customization of target vocabulary. This flexibility may support future research and pedagogical efforts in identifying optimal levels and types of input variability for phonetic learning. Notably, the AI voices used here were selected through ordinary auditory judgment, without any acoustic screening, yet still yielded learning outcomes at least comparable to human-recorded input. Because classroom teachers cannot realistically perform

acoustic analyses when selecting synthetic voices, this finding provides preliminary evidence that AI-generated voices can function as effective training input even when chosen perceptually. This speaks to the practical usability of perceptually selected AI voices, rather than to whether such voices necessarily reproduce the acoustic distinctions between AI and human speech reported in prior literature. The present findings contribute to the field by providing empirical support for the feasibility of utilizing AI-generated speech stimuli, while also underscoring the importance of grounding technological innovation in established theories of speech perception and L2 acquisition.

6. Conclusion

The present study investigated whether AI-generated speech can serve as an effective medium for classroom-based HVPT-informed vowel training and whether its effects differ from those of human-recorded speech. Both Human-based and AI-based training yielded robust improvements in overall vowel phoneme perception and AI-based training produced greater overall gains than human-based training. Generalization to new voices and untrained lexical items occurred in both groups, and the relative hierarchy of vowel-pair improvement remained stable across training modalities.

These findings indicate that AI-generated speech strengthens phonetic learning within an existing perceptual architecture rather than reorganizing it. Within the Speech Learning Model framework, AI input engages category-level restructuring comparable to that of human speech, indicating functional comparability in perceptual learning mechanisms. Theoretically, it was found that the observed training effectiveness from AI input does not require a change in the underlying acquisition mechanism. Pedagogically, the results support the feasibility of HVPT-informed vowel training as a scalable and accessible tool for L2 pronunciation training in a classroom context.

Two caveats should accompany these findings. First, the acoustic characteristics of the AI-generated and human-recorded stimuli were not independently verified; the present study drew on the platform's voice descriptions and on existing comparative literature (see Section 2.4) to characterize their likely differences. Direct acoustic measurement, particularly token-level variability analyses (e.g., F1 and F2 distributions, pitch range, speaking rate, and vowel space), would yield a more compelling account of why AI-based training produced greater gains in some conditions. The variability-based interpretation of the AI advantage offered in the Discussion should therefore be regarded as a hypothesis to be tested directly in future acoustic work. Second, perception was measured using a 2AFC identification task, which limits the granularity of inference because improvement on a given contrast cannot be decomposed into learning about each individual phoneme; future research could complement 2AFC with more fine-grained measures such as open-set identification, phoneme-level transcription, or multiple-choice identification with a larger response set.

Several extensions could further enrich this line of research. Future studies should systematically manipulate the number of speakers within both AI-generated and human-recorded conditions by comparing low-variability and high-variability versions of each training type. Such a design would make it possible to examine not only the role of variability as a core component of HVPT, but also whether AI-generated speech can provide a level of phonetic variability comparable to that of human speech. Future research should also examine whether comparable effects emerge across different age groups and other L2 phonological targets, and whether AI-based training facilitates production gains that are retained over time, as assessed through delayed post-tests. Moreover, future classroom-based pronunciation studies should include generalization tasks that examine whether training effects extend to sentence- and discourse-level speech. By incorporating assessment tasks that more closely approximate

real communicative contexts, future research could provide findings with greater pedagogical validity. Taken together, the present study demonstrates that AI-generated speech can function as a viable and accessible alternative to human-recorded input within classroom-based HVPT, offering an empirically grounded foundation for future research and pedagogy in L2 pronunciation instruction.

References

Al-Shami, F. and W. Cardoso. 2025. Text-to-speech in high-variability phonetic training: Focus on L2 phonological awareness. *Computer-Assisted Language Learning Electronic Journal* 26(6), 21-42.

Audacity Team. 2025. *Audacity* (version 3.7.5) [Computer software]. Available online at https://www.audacityteam.org

Baayen, R. H., D. J. Davidson and D. M. Bates. 2008. Mixed-effects modelling with crossed random effects for subjects and items. *Journal of Memory and Language* 59(4), 390-412.

Barrington, S., E. A. Cooper and H. Farid. 2025. People are poorly equipped to detect AI-powered voice clones. *Scientific Reports* 15, 11004.

Bates, D., M. Mächler, B. Bolker and S. Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67(1), 1-48.

Best, C. T. and M. D. Tyler. 2007. Nonnative and second-language speech perception: Commonalities and complementarities. In O.-S. Bohn and M. J. Munro, eds., *Language Experience in Second Language Speech Learning: In Honor of James Emil Flege,* 13-34. Amsterdam: John Benjamins.

Bione, T. and W. Cardoso. 2020. Synthetic voices in the foreign language context. *Language Learning and Technology* 24(1), 169-186.

Boersma, P. and D. Weenink. 2022. *Praat: Doing phonetics by computer* (version 6.2.09) [Computer software]. Available online at http://www.praat.org

Bohn, O.-S. and M. J. Munro. 2007. *Language Experience in Second Language Speech Learning: In Honor of James Emil Flege.* Amsterdam: John Benjamins.

Bradlow, A. R., R. Akahane-Yamada, D. B. Pisoni and Y. Tohkura. 1999. Training Japanese listeners to identify English /r/ and /l/: Long-term retention of learning in perception and production. *Perception and Psychophysics* 61(5), 977-985.

Bradlow, A. R., D. B. Pisoni, R. Akahane-Yamada and Y. Tohkura. 1997. Training Japanese listeners to identify English /r/ and /l/: IV. Some effects of perceptual learning on speech production. *Journal of the Acoustical Society of America* 101(4), 2299-2310.

Brekelmans, G., B. G. Evans and E. Wonnacott. 2025. Training child learners on nonnative vowel contrasts with phonetic training: The role of task and variability. *Language Learning* 75(3), 666-701.

Carlet, A. and J. Cebrian. 2014. Training Catalan speakers to identify L2 consonants and vowels: A short-term high variability training study. *Concordia Working Papers in Applied Linguistics* 5, 85-98.

Cheng, B., X. Zhang, S. Fan and Y. Zhang. 2019. The role of temporal acoustic exaggeration in high variability phonetic training: A behavioral and ERP study. *Frontiers in Psychology* 10(1178), 1-28.

Choe, S., H. Park and H. Ahn. 2025. The efficacy of high variability phonetic training for L2 speech perception in EFL contexts: A meta-analytic approach. *Korean Journal of English Language and Linguistics* 25, 1416-1443.

Davies, M. 2008-. *The Corpus of Contemporary American English (COCA).* Available online at https://www.english-corpora.org/coca

Flege, J. E. 1995. Second language speech learning: Theory, findings, and problems. In W. Strange, ed., *Speech Perception and Linguistic Experience: Issues in Cross-Language Research*, 233-277. Timonium, MD: York Press.

Flege, J. E. and O.-S. Bohn. 2021. The revised Speech Learning Model (SLM-r). In R. Wayland, ed., *Second Language Speech Learning: Theoretical and Empirical Progress*, 3-83. Cambridge: Cambridge University Press.

Flege, J. E., O.-S. Bohn and S. Jang. 1997. Effects of experience on non-native speakers' production and perception of English vowels. *Journal of Phonetics* 25(4), 437-470.

Fox, J. and S. Weisberg. 2019. *An R Companion to Applied Regression*. 3rd ed. Thousand Oaks, CA: Sage.

Hong, S. 2007. The characteristics of vowel identification errors of university-level Korean students of American English: HCA. *Language and Linguistics* 39, 257-277.

Hong, S. 2012. The relative perceptual easiness between perceptually assimilated vowels for university-level Korean learners of American English and measurement bias in an identification test. *Studies in Phonetics, Phonology and Morphology* 18(3), 491-511.

Hwang, H. and H. Y. Lee. 2015. The effect of high variability phonetic training on the production of English vowels and consonants. In *Proceedings of 18th International Congress of Phonetic Sciences*, 1041-1045.

Ingram, J. C. L. and S.-G. Park. 1997. Cross-language vowel perception and production by Japanese and Korean learners of English. *Journal of Phonetics* 25(3), 343-370.

John, P. and W. Cardoso. 2016. A comparative study of text-to-speech and native speaker output. In *Proceedings of the 4th Annual Meeting on Language Teaching*, 78-96.

Kahoot! 2025. *Kahoot!* [Game-based learning platform]. Available online at https://kahoot.com

Kim, J.-E. 2010. Perception and production of English front vowels by Korean speakers. *Phonetics and Speech Sciences* 2(1), 51-58.

Kim, Y. and C. Seong. 2025. Detection of AI-generated speech using acoustic parameters. *Phonetics and Speech Sciences* 17(3), 15-22.

Lee, A. H. and R. Lyster. 2017. Can corrective feedback on perception affect production? *Applied Psycholinguistics* 38(2), 371-393.

Lee, B., S. Lee, C. Kim, M. Ko and S. Kim. 2015. *Middle School English 3*. Seoul: Dong-A Publishing.

Lee, B., L. Plonsky and K. Saito. 2020. The effects of perception-vs. production-based pronunciation instruction. *System* 88, 102185.

Lee, H. Y. and H. Hwang. 2016. Gradient of learnability in teaching English pronunciation to Korean learners. *Journal of the Acoustical Society of America* 139, 1859-1872.

Lee, K. and M. Cho. 2015. Perception of English vowels by Korean learners: Comparisons between new and similar L2 vowel categories. *The Journal of the Korea Contents Association* 15(8), 579-587.

Lee, S. and H. Baek. 2025. The role of perceptual and acoustic similarity in learners' perception of L2 vowels. *Korean Journal of English Language and Linguistics* 25, 1084-1101.

Lee, S. and M. Cho. 2018. Predicting L2 vowel identification accuracy from cross-language mappings between L2 English and L1 Korean. *Language Sciences* 66, 183-198.

Lee, S. A. S. and G. K. Iverson. 2012. Vowel category formation in Korean-English bilingual children. *Journal of Speech, Language, and Hearing Research* 55(5), 1449-1462.

Lenth, R. and J. Piaskowski. 2026. *emmeans: Estimated Marginal Means, aka Least-Squares Means* (version 2.0.2) [R package]. Available online at https://rvlenth.github.io/emmeans

Leong, C. X. R., J. M. Price, N. J. Pitchford and W. J. B. van Heuven. 2018. High variability phonetic training in adaptive adverse conditions is rapid, effective, and sustained. *PLoS ONE* 13(10), e0204888.

Li, Y. 2024. A comparison of perception-based and production-based training approaches to adults' learning of L2 sounds. *Language Learning and Development* 20(3), 232-248.

Liakin, D., W. Cardoso and N. Liakina. 2017. The pedagogical use of mobile speech synthesis (TTS): Focus on French liaison. *Computer Assisted Language Learning* 30(3-4), 325-342.

Lively, S. E., J. S. Logan and D. B. Pisoni. 1993. Training Japanese listeners to identify English /r/ and /l/: The role of phonetic environment and talker variability. *Journal of the Acoustical Society of America* 94(3), 1242-1255.

Logan, J. S., S. E. Lively and D. B. Pisoni. 1991. Training Japanese listeners to identify English /r/ and /l/: A first report. *Journal of the Acoustical Society of America* 89(2), 874-886.

LOVO. 2023. *Genny* [AI voice generator and text-to-speech platform]. Available online at https://lovo.ai

Mohammadkarimi, E. 2024. Exploring the use of artificial intelligence in promoting English language pronunciation skills. *LLT Journal: A Journal on Language and Language Teaching* 27(1), 98-115.

Moodle. 2025. *Moodle* [Learning management system]. Available online at https://moodle.org

Nagle, C. L. and M. M. Baese-Berk. 2022. Advancing the state of the art in L2 speech perception-production research: Revisiting theoretical assumptions and methodological practices. *Studies in Second Language Acquisition* 44(2), 580-605.

Park, M. and J. Lee. 2024. Korean ESL learners' production of English vowel contrasts: Developmental variations in L2 sound learning. *Korean Journal of English Language and Linguistics* 24, 1318-1332.

Piske, T. 2007. Implications of James E. Flege's research for the foreign language classroom. In O.-S. Bohn and M. J. Munro, eds., *Language Experience in Second Language Speech Learning: In Honor of James Emil Flege*, 301-314. Amsterdam: John Benjamins.

R Core Team. 2025. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing [Computer software]. Available online at https://www.R-project.org

Ren, Y., X. Tan, T. Qin, Z. Zhao and T.-Y. Liu. 2022. Revisiting over-smoothness in text to speech. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, 8197-8213.

Sakai, M. and C. Moorman. 2018. Can perception training improve the production of second language phonemes? A meta-analytic review of 25 years of perception training research. *Applied Psycholinguistics* 39(1), 187-224.

Shen, J., R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang and Q. V. Le. 2018. Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions. In *Proceedings of 2018 IEEE International Conference on Acoustics, Speech, and Signal Processing*, 4779-4783.

Smith, G., W. Cardoso and C. García Fuentes. 2016. Text-to-speech synthesizers: Are they ready for the second language classroom? In *Proceedings of the 4th Annual Meeting on Language Teaching*, 97-111.

Tan, X., T. Qin, F. Soong and T.-Y. Liu. 2021. A survey on neural speech synthesis. *arXiv preprint* arXiv:2106.15561.

Thomson, R. I. 2012. Improving L2 listeners' perception of English vowels: A computer-mediated approach. *Language Learning* 62(4), 1231-1258.

Thomson, R. I. 2018. High variability [pronunciation] training (HVPT). A proven technique about which every language teacher and learner ought to know. *Journal of Second Language Pronunciation* 4(2), 208-231.

Tsukada, K., D. Birdsong, E. Bialystok, M. Mack, H. Sung and J. E. Flege. 2005. A developmental study of English vowel production and perception by native Korean adults and children. *Journal of Phonetics* 33(3), 263-290.

Tyler, M. D. 2019. PAM-L2 and phonological category acquisition in the foreign language classroom. In A. M. Nyvad, M. Hejná, A. Højen, A. Jespersen, M. Hjortshøj and M. Sørensen, eds., *A Sound Approach to Language Matters - In Honor of Ocke-Schwen Bohn*, 607-630. Aarhus, Denmark: Aarhus University Press.

Uchihara, T., M. Karas and R. I. Thomson. 2025. High variability phonetic training (HVPT): A meta-analysis of L2 perceptual training studies. *Studies in Second Language Acquisition* 47, 794-827.

Vancova, H. 2023. AI and AI-powered tools for pronunciation training. *Journal of Language and Cultural Education* 11(3), 12-24.

Wong, J. W. S. 2026. What do the stimuli in high variability phonetic training tell us about second language perception? Current findings, implications, and the way forward. In M. Reed and J. M. Levis, eds., *The Handbook of Second Language Listening*, 155-168. Hoboken, NJ: John Wiley & Sons.

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

Yang, B. 2010. College students' production and perception of English vowels. *English Language Teaching* 22(4), 165-184.

Zhang, X., B. Cheng, D. Qin and Y. Zhang. 2021. Is talker variability a critical component of effective phonetic training for nonnative speech? *Journal of Phonetics* 87, 101071.

Examples in: English
Applicable Languages: English
Applicable Level: Tertiary

Appendix A. Descriptive Statistics for Phoneme Identification Accuracy

Table A1. Descriptive Statistics by Group, Time, and Voice Type

Group	Time	Voice Type	*M* (% accuracy)	*SD*	*n*
Human	Pre	New	75.1	12.3	40
		Familiar	74.7	12.3	40
	Post	New	81.4	9.3	40
		Familiar	83.8	10.2	40
AI	Pre	New	73.0	13.3	34
		Familiar	72.1	14.2	34
	Post	New	85.0	10.5	34
		Familiar	83.7	7.49	34

Note. M and SD are reported in percentage points based on participant-level mean accuracy scores.

Table A2. Descriptive Statistics by Group, Time, and Word Type

Group	Time	Word Type	*M* (% accuracy)	*SD*	*n*
Human	Pre	untrained	75.3	11.2	40
		trained	74.5	12.8	40
	Post	untrained	80.5	9.55	40
		trained	84.6	9.54	40
AI	Pre	untrained	73.8	12.0	34
		trained	71.3	15.8	34
	Post	untrained	84.3	7.4	34
		trained	84.4	10.9	34

Note. M and SD are reported in percentage points based on participant-level mean accuracy scores.

Table A3. Descriptive Statistics by Group, Time, and Vowel Contrast

Group	Time	*Contrast*	*M* (% accuracy)	*SD*	*n*
Human	Pre	/i-ɪ/	80.3	17.7	40
		/ɛ-æ/	51.2	23.6	40
		/ɛ-eɪ/	98.4	4.19	40
		/ɑ-ʌ/	71.2	19.4	40
		/ɔ-oʊ/	80.9	19.6	40
		/u-ʊ/	67.2	18.5	40
	Post	/i-ɪ/	81.2	19.6	40
		/ɛ-æ/	80.6	19.6	40
		/ɛ-eɪ/	98.4	4.19	40
		/ɑ-ʌ/	60.9	19.4	40
		/ɔ-oʊ/	91.9	11.5	40
		/u-ʊ/	82.2	14.7	40
AI	Pre	/i-ɪ/	78.7	17.0	34
		/ɛ-æ/	50.0	27.7	34
		/ɛ-eɪ/	97.4	7.4	34
		/ɑ-ʌ/	61.8	19.4	34
		/ɔ-oʊ/	79.4	19.9	34
		/u-ʊ/	68.0	20.0	34
	Post	/i-ɪ/	79.4	19.2	34
		/ɛ-æ/	87.9	14.3	34
		/ɛ-eɪ/	98.9	3.6	34
		/ɑ-ʌ/	62.1	18.6	34
		/ɔ-oʊ/	90.8	15.2	34
		/u-ʊ/	87.1	13.2	34

Note. M and SD are reported in percentage points based on participant-level mean accuracy scores.