World Journal of Educational Research
ISSN 2375-9771 (Print) ISSN 2333-5998 (Online)
Vol. 13, No. 5, 2026
www.scholink.org/ojs/index.php/wjer

Original Paper

Using TTS to Develop Phonetic Training Materials

Xiaoting Zhu[1]

[1] College of Humanities and Law (College of Foreign Languages), Zhejiang A&F University, Hangzhou, Zhejiang Province, China

Received: June 29, 2026 Accepted: August 22, 2026 Online Published: September 18, 2026

doi:10.22158/wjer.v13n5p1 URL: http://dx.doi.org/10.22158/wjer.v13n5p1

Abstract

This study investigated whether TTS-generated speech can serve as effective training material in High Variability Phonetic Training (HVPT) and explored the learning patterns reflected in the perceptual results. A group of native Chinese undergraduates completed six sessions of self-paced, take-home HVPT for English vowels. Their vowel perception performance was measured with identification (ID) and discrimination (DISC) tasks before and after training. The results showed that accuracy on both tasks at posttest was significantly higher than at pretest, and the training did not have interactions with vowels or vowel contrasts in either task, suggesting that TTS-generated material is effective and that the training effect is not specific to certain vowels or vowel contrasts. Moreover, learners with lower initial ID performance benefited more from training, whereas the improvement in the DISC task did not depend on initial performance. These findings indicate that TTS-generated speech can serve as effective HVPT material for L2 vowel perception, with benefits that vary with learners' initial performance and task type.

Keywords

text-to-speech (TTS), High Variability Phonetic Training (HVPT), Second Language (L2) vowel perception

1. Introduction

Acquiring sounds in a second language (L2) remains a challenge, especially for adult learners; however, phonetic training has been shown to facilitate L2 perception and learning. High Variability Phonetic Training (HVPT), a well-established phonetic training paradigm, is characterized by exposure to natural variability (i.e., using multiple talkers, multiple phonetic contexts) and immediate feedback that indicate whether participants' answer was correct or not (Iverson et al., 2012; Logan et al., 1991). It has been shown to produce long-lasting generalization to novel speakers and phonetic contexts (Iverson & Evans, 2009).

As a proven paradigm, HVPT has not yet been widely adopted in practice. One reason is the complexity of preparing the training material. The training material is a complex system that involves a huge number of systematically varied tokens recorded by human speakers and screened for accuracy (Thomson, 2018). This process is thus laborious, and the material is difficult to modify or expand once set. In this sense, recent advances in neural text-to-speech (TTS) offer a potential alternative. Neural TTS is able to generate highly natural and controllable synthetic voices and allows speech materials to be produced more readily and flexibly. For example, users can adjust pitch, rate, accent, and timbre, customize the input information, and regenerate the speech. Moreover, TTS-generated speech is able to provide human-comparable fidelity and comprehensibility. For example, John and Cardoso (2016) rated that contemporary TTS segments as indistinguishable from a native speaker in terms of problematic English consonants, such as interdental and glottal fricatives, lateral and rhotic liquids, and for the tense-lax vowel contrasts /i-ɪ/, /u-ʊ/, and /ɛ-æ/. Bione and Cardoso (2020) further found that TTS and human voices were perceptually comparable in an identification task of the past tense -ed.

Despite these advances, TTS systems exhibit a tendency toward spectral smoothing, known as over-smoothness. They tend to generate mel-spectrograms that approximate the average of training distributions, thus presenting somewhat lower acoustic variation indices compared to human speech (Ren et al., 2020). Human speech, by contrast, contains wide micro-variability, originating from articulation and coarticulation, as well as speaker-specific anatomical differences, yielding broader within- and between-speaker acoustic distributions. Variability is particularly thought to be critical in HVPT, because it teaches learners the acoustic cues that are most robust across different situations, while a limited set of training material leads to learning that is specific to the trained stimuli (Lively et al., 1993; Logan et al., 1991). Therefore, whether TTS-generated speech can provide sufficiently useful input for perceptual learning in HVPT remains a question.

Although this over-smoothness tendency seems to undermine the key characteristic of HVPT, a few studies have begun to examine TTS-generated speech as training material in L2 speech learning and have suggested that such input is effective. For example, Al-Shami and Cardoso (2025) integrated TTS voices in HVPT training in a semi-autonomous environment (beyond the classroom), and the results revealed significant improvement in phonological awareness of English past tense allomorphs. Even more relevant, Park and Ahn (2026) compared training gains between TTS-generated and human-recorded material using an identification task in structured classroom activities, and found that the TTS group had greater overall gains. However, because vowel perception was assessed by a 2-alternative forced-choice (2AFC) identification task, the results tell little about how the improvement occurred, such as at which level of phonetic processing.

The present study trained L1 Chinese speakers on English vowels using TTS-generated material in a self-paced, take-home way to evaluate whether TTS-generated speech as phonetic training material can improve vowel perceptual performance and, if so, to explore the patterns of learning reflected in the resulting perceptual changes. The analysis examined not only whether perceptual performance

improved overall, but also how the effect was distributed across vowels and vowel contrasts, and learners with different initial levels of performance. Two kinds of perceptual tasks were used to examine the training effect, namely identification (ID) and discrimination (DIS) tasks. An ID task requires participants to identify or label a sound stimulus, while a DIS task requires participants to judge whether two or more sounds are in the same category. The two tasks are thought to measure different aspects of L2 perception (i.e., generalization to novel speakers and words, low-level auditory processing) from the training.

2. Method

2.1 Participants

A total of 13 undergraduate students (3 males and 10 females) participated in the study, aged between 18 and 19. Nine additional subjects began the training but did not complete it; one was excluded because the categorical discrimination was 97% correct in the pretest. All participants were native Chinese speakers with the College English Test Band 4 (CET-4) passed, and none of them self-reported that they had any known hearing or language problems. Informed consent was obtained from all participants, who also received a small incentive for their participation.

2.2 Apparatus and Stimuli

The pre- and posttests were conducted through a self-developed website on participants' mobile phones in a quiet room. All training sessions were conducted by participants on their own mobile phones. The recordings used in the pre- and posttests were made in a quiet room.

The pre- and posttest stimuli comprised recordings of English /b/-V-/t/ words from a male native speaker of British English. None of these words were used in the training corpus, such that all pre- and posttests measured generalization to new stimuli. The speaker read the words beat /i/, bit /ɪ/, bet /ɛ/, bat /a/, Bart /ɑ/, but /ʌ/, bot /ɒ/, bought /ɔ/, bout /aʊ/, and boat /əʊ/ several times in a natural way. These vowels were selected because they (1) created real English words, (2) made vowel contrasts, (3) were comparatively difficult for Chinese students to learn. For example, /ʊ/ did not make a real English word in strict /b/-V-/t/ form, thus was excluded, while /u/ created a real English word, boot, but the vowel mainly made a contrast with the /ʊ/ sound for Chinese learners (Xue, 2018), so that /u/ was excluded as well. Nonetheless, /ɜ/ did create a real word, Bert, but it seldom contrasted with other vowels for Chinese learners; thus, /ɜ/ was excluded.

The training stimuli were created by 6 TTS voices from Speechify (3 males and 3 females, all in British accent). The words were minimal pairs grouped by dividing 10 British English vowels: /i/, /ɪ/ (e.g., eel, ill); /ɛ/, /ɑ/, /a/, /ʌ/ (e.g., pet, part, pat, putt); and /ɒ/, /ɔ/, /aʊ/, /əʊ/ (e.g., cod, called, cowed, code). These vowels were clustered because they were mutually confusable by many listeners based on the results of identification and discrimination tests in a previous relevant study (Xue, 2018). There were 5 sets of minimal pair words for each of the clusters, for a total of 50 words.

2.3 Procedure

(1) Pre-/post-tests

The pre- or posttests were carried out one day before or after the training sessions. Participants first completed the identification (ID) test and then the discrimination (DISC) test.

For ID test, participants heard one English /b/-V-/t/ word on each trial and were asked to match the word in a closed-set response with all 10 words as options presented on a mobile screen. For stimuli that might have been unfamiliar to participants, a more common English word containing the same vowel was provided alongside the target word, for example, bout (out). They received no feedback and were not able to replay the stimulus. Before the formal test, participants completed an initial 10-trial practice session to familiarize themselves with the procedure and could ask any questions about the task instructions. For the formal test, there were four repetitions of the 10 vowels for a total of 40 trials presented in a random order.

For DISC test, participants heard the sound of three English /b/-V-/t/ words on each trial without words being presented. Two words were the same, and one was different. Students were asked to decide the different one by tapping the corresponding option. Again, they received no feedback and were able to play the stimulus only once. Participants were given an initial 6-trial practice session to familiarize themselves with the procedure and could ask any questions regarding the task instructions. As in the formal test, there were six pairs of words, with each played eight times, for a total of 48 trials. For example, in the pair /i/–/ɪ/, four trials with /i/ and four trials with /ɪ/ were the odd stimulus, with the odd stimulus played first, second, or third randomly. All the trials were also played in a random order. The vowel pairs were /i/–/ɪ/, /a/–/ʌ/, /ɑ/–/a/, /ʌ/–/ɑ/, /ɒ/–/ɔ/, /əʊ/–/aʊ/. The pairs were selected based on English vowel identification and discrimination results in a previous relevant study. (Xue, 2018).

(2) Training

Training had six sessions, each of them typically lasting about 30 minutes. The entire course of training was designed as a take-home, self-paced listening assignment, but it was still required to be completed over no fewer than 6 days (no more than one session a day) and no more than 14 days. There was a different talker for each session.

On each trial, participants heard a stimulus word only once and tapped one option from the given 2 or 4 closed-set alternatives (depending on how vowels were grouped, as described in the stimuli part). The sound of the stimulus word was played before the options were shown with the intention to avoid priming effect. For stimuli that might have been unfamiliar to participants, a more common English word containing the same vowel was provided alongside the target word, for example, cat, cart, cut, ket (get).

There was also feedback. If participants made a correct response, the chosen option turned green, and the sound was played one more time. If they made a wrong response, the chosen word turned red, the correct word turned green, and the sound played. Then the correct and chosen word played in turn twice, so that students could learn the contrasts between the two words.

In each session, there was a 6-trial initial practice for participants to get familiar with the TTS voice and the training process. The formal training had 140 trials. The initial and last 50 trials went through all 50 stimulus words in random order. In between, the two vowel clusters that had the lowest accuracy in the initial block would go through another 40 trials, so that training could be customized to some extent.

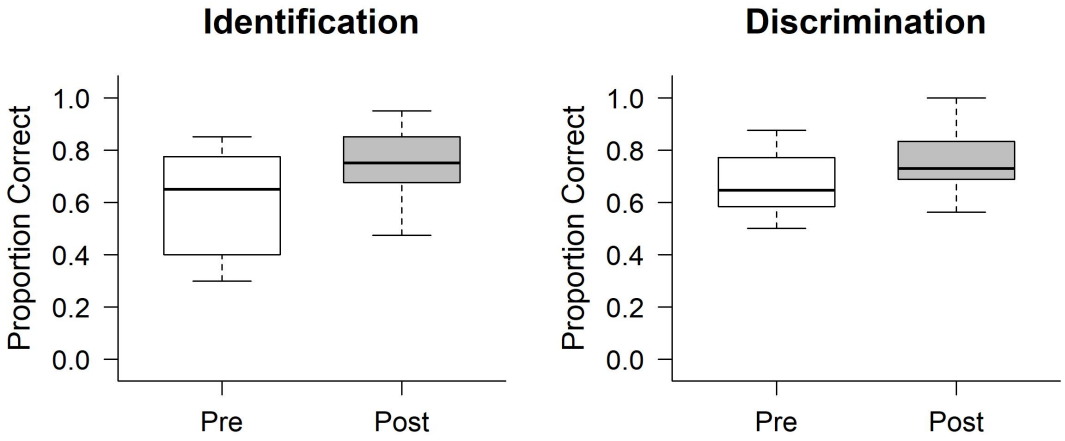

Figure 1. Boxplots of Identification and Discrimination Accuracy at Pretest and Posttest

3. Result

3.1 Identification (ID)

Figure 1 includes the identification accuracy before and after training, while Figure 2 presents the pre- and posttest identification accuracy for each vowel. A 2 X 10 repeated-measures ANOVA on arc-sin transformed scores showed a significant main effect of training, $F(1,12) = 25.77$, $p < .001$, $\eta p^2 = 0.68$. Mean ID accuracy increased from 59% at pretest to 74% at posttest. There was also a significant main effect of vowels, $F(9,108) = 14.45$, $p < .001$, $\eta p^2 = 0.55$, suggesting that some vowels were more difficult than others to identify. However, the non-significant interaction between training and vowel, $F(9,108) = 0.29$, $p = 0.977$, $\eta p^2 = 0.02$, demonstrated that although some vowels seemed to be more difficult, training improved ID accuracy to a similar extent across vowels.

For individual differences in identification tests, there was a significant negative Pearson correlation ($r = -0.67$, $p = 0.013$) between pretest ID accuracy and ID improvement; thus, participants with lower initial performance tended to benefit more from the training.

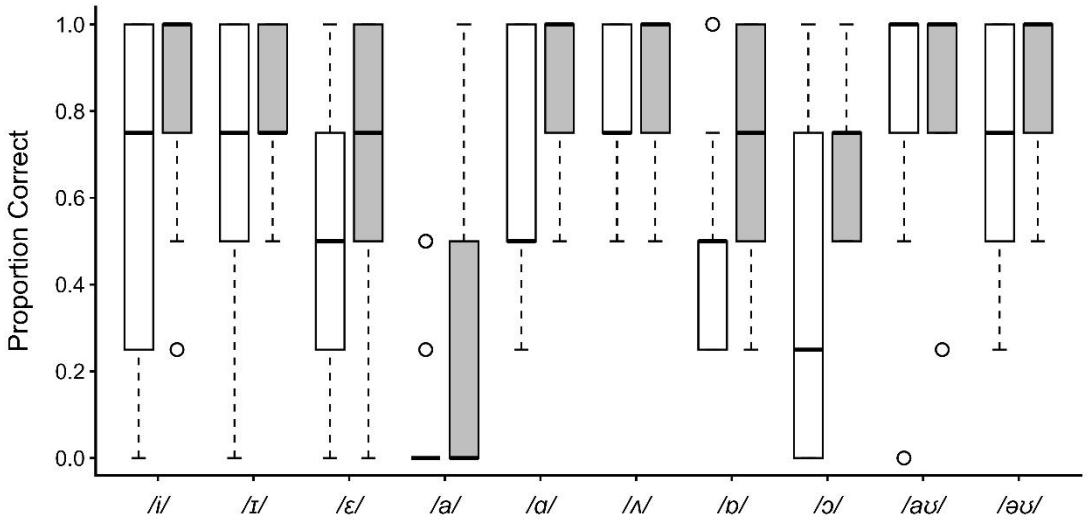

Figure 2. Boxplots of Identification Accuracy for each Vowel at Pretest (white boxes) and Posttest (gray boxes)

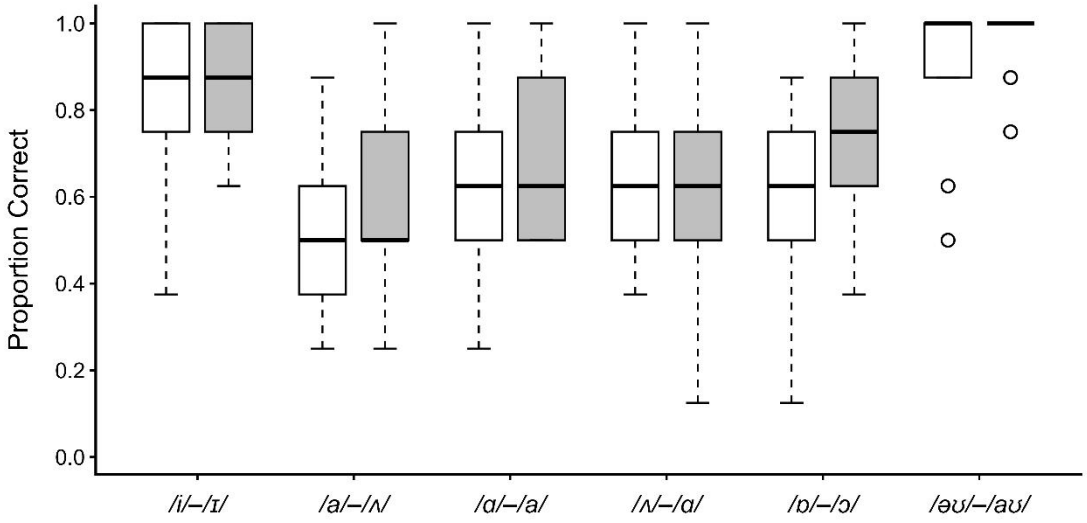

Figure 3. Boxplots of Identification Accuracy for each Vowel Pair at Pretest (white boxes) and posttest (gray boxes)

3.2 Discrimination

Figure 1 includes the discrimination accuracy before and after training, while Figure 3 presents the pre- and posttest identification accuracy for each vowel contrast. A 2 X 6 repeated-measures ANOVA on arc-sin transformed scores showed a significant main effect of training, $F(1,12) = 7.56$, $p = 0.018$, $\eta p^2 = 0.39$. Mean DISC accuracy increased from 69% at pretest to 75% at posttest. There was also a significant main effect of vowels, $F(5,60) = 19.90$, $p < .001$, $\eta p^2 = 0.62$, suggesting that some vowel contrasts were more or less difficult than others to discriminate. However, again, the non-significant

interaction between the training and vowel, $F(5,60) = 0.75$, $p = 0.588$, $\eta p^2 = 0.06$, demonstrated that the training effect was relatively general across all the vowel contrasts, despite the differences in their overall accuracy.

However, for individual differences, a non-significant correlation between pretest performance and improvement ($r = -0.41$, $p = 0.160$) did not provide sufficient evidence to suggest that the training effect was dependent on participants' initial performance.

3.3 Relation between Identification and Discrimination

Pearson correlation analyses revealed that categorical discrimination and identification accuracy were not significantly correlated at either pretest ($r = 0.25$, $p = 0.42$) or posttest ($r = 0.36$, $p = 0.23$). Similarly, the improvement in categorical discrimination (i.e., posttest accuracy minus pretest accuracy) was not significantly correlated with the improvement in identification accuracy ($r = -0.24$, $p = 0.436$). These results indicated that the ability to identify and discriminate in these tasks did not seem to be directly correlated, although training brought their performance numerically closer.

4. Discussion

This study investigated the effectiveness of TTS-generated material in HVPT on Chinese learners' perception of English vowels and explored the patterns of learning reflected in the resulting perceptual changes. The results showed that TTS-generated training material could help students improve significantly in both ID and DISC tasks. This overall result was in line with a previous study using TTS-generated input as training material in a 2AFC ID task (Park & Ahn, 2026). Besides, at the item level, even though some of the vowels or vowel contrasts were significantly more or less difficult than others, the training produced a relatively general effect across them rather than being specific to certain vowels or vowel contrasts. The current study thus adds more evidence to support that TTS-generated training material is generally effective in improving English vowel perception.

These results also offer some insight into the ongoing discussion on the roles of different characteristics in HVPT. Traditional HVPT has emphasized exposure to natural variable speech because variability is assumed to help learners attend to the most robust acoustic cues (Lively et al., 1993; Logan et al., 1991), whereas TTS-generated speech has been shown to exhibit reduced acoustic variability compared with natural speech (Ren et al., 2022). Nevertheless, the current training elicited learning similar to that in other HVPT studies, suggesting that this reduced variability is sufficient to support learning. That being said, other factors in HVPT also contribute to its effectiveness, such as immediate feedback during training (Logan et al., 1991). For example, Iverson et al. (2012) found that exposure to greater variability alone did not seem to contribute to greater learning in HVPT, suggesting that variability alone does not automatically guarantee that attention is directed to critical phonetic differences. In this sense, the current training still helps direct learners' attention to relevant phonetic differences through task design (e.g., immediate feedback during training), which contributes to its effectiveness.

More comparisons in pre- and posttest help further analyze the learning outcomes of this training. The current study did not find a significant correlation between ID and the DISC task at either the pretest or posttest, although numerically higher at the posttest. This pattern seems to differ from previous studies (e.g., Gong et al., 2019; Iverson et al., 2012), which used human-recorded speech and often found positive correlations between ID and DISC tasks. These two tasks are often thought to be correlated as they share many of the same abilities, particularly when a discrimination task requires participants to make categorical judgement (Cebrian et al., 2024). However, successful discrimination performance may not always reflect the use of distinct language-specific phonetic codes; at least in some paradigms (e.g., the AXB paradigm), discrimination judgement is possibly made on the basis of low-level, perhaps auditory, memory codes (Højen & Flege, 2006). Although the present study employed an oddity paradigm rather than an AXB one, the absence of significant correlations between ID and DISC in the present study may still reflect task-specific demands placed on subjects.

At the same time, the current study found a strong negative correlation between pretest ID performance and ID improvement, whereas no such significant correlation was observed for the DISC task. One could have questioned whether this improvement stems from increased familiarity with orthographic labels. However, for unfamiliar words, a more common word was provided next to it, such as bout (out), so the problem of orthographic labels does not seem to count much for this pattern. A more plausible explanation could be that participants with lower initial performance had no yet established stable phonetic categories and therefore relied more heavily on auditory rather than phonetic codes in the DISC task at pretest, and had greater difficulty in mapping the heard stimuli to phonetic category. The numerically higher correlation, although still not significant, between ID and the DISC task at posttest is broadly consistent with this interpretation, as improved identification may have brought performance on the two tasks into closer alignment. In this sense, the present findings may be more consistent with the idea that training stabilizes existing L2 vowel categories than restructuring them (eg., Gong et al., 2019; Iverson & Evans, 2009; Iverson et al., 2012).

5. Conclusion

The current study investigated whether TTS-generated material can be effective for HVPT training in a self-paced and take-home manner, and analyzed the pattern of learning outcomes. The training improved both ID and DISC performance, with the training effect being relatively general across vowels and vowel contrasts. For individual differences, participants with lower initial ID performance benefited more from training, whereas the improvement of DISC task did not seem to be dependent on initial performance. Combined with the absence of correlations between ID and DISC tasks at pretest and posttest, as well as between their improvement, these findings suggest that the training has strengthened or stabilized the existing vowel categories.

The limitations of current study should also be acknowledged. First, the study focused on several representative vowel contrasts. While these vowel contrasts were identified as challenging in previous

studies, a broader range of vowel contrasts might reveal more detailed underlying perceptual process. Second, the TTS-generated training materials were not acoustically analyzed or perceptually validated. Future studies could involve such analysis to evaluate the related features of TTS-generated speech, and made clearer its suitability as training materials in HVPT.

Acknowledgement

This work was supported by Grant No. JG2025065 from Zhejiang A&F University Teaching Reform Special Project.

References

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

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

Cebrian, J., Gavaldà, N., Gorba, C., & Carlet, A. (2024). Differential effects of identification and discrimination training tasks on L2 vowel identification and discrimination. *Studies in Second Language Acquisition*, *46*(4), 1-25. https://doi.org/10.1017/S0272263124000408

Gong, J., Yang, Z., Ji, X., & Wang, F. (2019). Investigating the effectiveness of auditory training on Chinese listeners' perception of English vowels. P*roceedings of the 19th International Congress of Phonetic Sciences*, 1530-1534.

Højen, A., & Flege, J. E. (2006). Early learners' discrimination of second-language vowels. *The Journal of the Acoustical Society of America*, *119*(5), 3072-3084. https://doi.org/10.1121/1.2184289

Iverson, P., & Evans, B. G. (2009). Learning English vowels with different first-language vowel systems II: Auditory training for native Spanish and German speakers. *The Journal of the Acoustical Society of America*, *126*(2), 866-877. https://doi.org/10.1121/1.3148196

Iverson, P., Pinet, M., & Evans, B. G. (2012). Auditory training for experienced and inexperienced second-language learners: Native French speakers learning English vowels. *Applied Psycholinguistics*, *33*(1), 145-164. https://doi.org/10.1017/S0142716411000300

John, P., & Cardoso, W. (2016). A comparative study of text-to-speech and native speaker output. In J. Demperio, E. Rosales, & S. Springer (Eds.), *Proceedings of the Meeting on English Language Teaching* (pp. 78-96). Université du Québec à Montréal Press.

Lively, S. E., Logan, J. S., & Pisoni, D. B. (1993). Training Japanese listeners to identify English /r/ and /l/. II: The role of phonetic environment and talker variability in learning new perceptual categories. *Journal of the Acoustical Society of America*, *94*, 1242-1255. https://doi.org/10.1121/1.408177

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

Park, H., & Ahn, H. (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.

Ren, Y., Tan, X., Qin, T., Zhao, Z., & Liu, T.-Y. (2022). Revisiting over-smoothness in text to speech. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers), 8197-8213.

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.

Xue, D. (2018). *An experimental study on Chinese EFL learners' perception and production of English vowels* (Master's thesis, Jiangsu University of Science and Technology).