pISSN 2005-8063
eISSN 2586-5854
2025. 9. 30.
Vol.17 No.3
pp. 41-49

말소리와 음성과학

Phonetics and Speech Sciences

한국음성학회지

https://doi.org/10.13064/KSSS.2025.17.3.041

The role of speaking rates in High Variability Phonetic Training*

Jieun Lee[1,**] · Hanyong Park[2]

[1]*Department of Linguistics, University of Kansas, Lawrence, KS, USA*
[2]*Department of Linguistics, University of Wisconsin-Milwaukee, Milwaukee, WI, USA*

Abstract

This study examined whether incorporating multiple speaking rates into High Variability Phonetic Training (HVPT) facilitates second language (L2) learners' acquisition of a non-native phonological contrast. Native English-speaking naïve learners of Korean were trained to identify the Korean three-way stop contrast, which differs from English stop voicing contrasts in the relative weighting of voice onset time (VOT) and fundamental frequency (f0). Participants were assigned to a Multi-group, trained with stimuli from multiple talkers at three speaking rates (slow, normal, fast), or a Single group, trained with the same talkers only at the slow rate. HVPT sessions spanned three days, followed by everyday identification tests, a new talker generalization test, and a retention test one week later. Test stimuli included only slow-rate items. Acoustic analysis showed that faster speaking rates increased VOT overlap between lenis and aspirated stops while f0 distinctions remained relatively stable. Preliminary results showed that the Multi-group consistently outperformed the Single group in training, generalization, and retention. We discuss speaking rate as a potentially meaningful dimension of stimulus variability in HVPT and highlight its promise for future large-scale research to examine the extent to which incorporating multiple speaking rates promotes more robust L2 perceptual learning.

Keywords: high variability phonetic training, speaking rate, L2 contrast learning

1. Introduction

The present study aims to develop an effective second language (L2) perceptual training method that helps non–native listeners use perceptual acoustic cues in a nativelike manner when identifying non–native phonological contrasts. To this end, the current study investigated the effectiveness of High Variability Phonetic Training (HVPT) that incorporated various speaking rates. The target case involved naïve English–speaking learners of Korean acquiring the Korean three–way stop consonant contrast. This contrast poses challenges for English listeners due to the reversed relationship between primary and secondary acoustic dimensions—Voice Onset Time (VOT) & fundamental frequency (f0)—in word–initial position (e.g., Lee & Park, 2024).

We examined whether increasing phonetic variability in L2 training stimuli by including multiple speaking rates (slow, normal, and fast) would enhance learners' ability to identify the target L2 contrast and promote greater generalization of learning compared to

* This paper is developed from an earlier paper presented at 22nd Mid-Continental Phonetics & Phonology Conference.

** jieunlee@ku.edu, Corresponding author

Received 14 August 2025; Revised 8 September 2025; Accepted 8 September 2025

a training paradigm using a single speaking rate (slow). To address this question, we conducted a three–day HVPT experiment using training stimuli recorded at three speaking rates and compared participants' performances across two training conditions (Single vs. Multi).

1.1. High Variability Phonetic Training (HVPT) with Speaking Rates

HVPT is a perceptual training method that presents learners with a wide range of acoustic variability, typically using multiple–talker produced stimuli, to promote robust L2 category formation. In L2 perceptual phonetic training, it has been widely investigated whether training with highly variable phonetic stimuli is more effective than training with low–variability stimuli, such as those including single talker materials (e.g., Bradlow et al., 1999; Lively et al., 1993; Logan et al., 1991). HVPT assumes that exposure to a wide range of L2 exemplars, similar to those encountered in real–world speech, helps learners build robust category representations. This, in turn, can lead to better generalization of learning, such as recognizing the target contrast in novel words or across unfamiliar talkers, as well as longer retention of training effect (Bradlow et al., 1999). High stimulus variability is often achieved by increasing the number of training talkers or varying phonetic environments. For example, Logan et al. (1991) demonstrated that Japanese learners of English trained with stimuli produced by multiple talkers and in various phonetic environments generalized their learning not only to new words produced by one of the training talkers but also to words produced by an unfamiliar talker. It was suggested that training stimuli variability helped learners form robust L2 English categories and overcome talker–specific learning by exposing them to the full range of acoustic–phonetic cues created by different phonetic environments.

In addition to talkers and phonetic environments, what other factors can serve as meaningful dimensions providing phonetic variability in HVPT? Hirata et al. (2007) tested three types of HVPT involving speaking rate variation: slow–only, fast–only, and slow–fast. They found that training with two speaking rates (i.e., slow–fast) resulted in higher accuracy in identifying the target L2 Japanese vowel length contrasts, compared to training with a single speaking rate. However, the slow–fast training effect was marginal, with overall improvement rates of 8.6% for slow–only training and 9.1% for slow–fast training. Although Hirata et al. (2007) supports the potential effectiveness of HVPT with speaking rate variability, Sonu et al. (2013) reported somewhat different results. They compared two groups of Korean learners of Japanese in identifying the Japanese length contrast (geminate vs. singleton). One group was trained with a single speaking rate, while the other group was trained with three different speaking rates. Sonu et al. (2013) hypothesized that high stimulus variability might help learners adapt their use of duration cues in identifying the target length contrast, which vary significantly depending on speaking rates. By exposing learners to multiple speaking rates, they were expected to rely on relative durational differences rather than absolute ones in perceiving the target contrast. The results, however, showed that both groups improved in identifying the target contrast, with no significant difference in accuracy between them.

Although both Hirata et al. (2007) and Sonu et al. (2013) targeted L2 contrasts for which a primary cue is considerably affected by speaking rates, their findings on the robustness of training effects with multiple speaking rates were mixed. To further investigate whether HVPT with speaking rates can reliably enhance L2 phonological contrast learning, the current study examines another case of L2 learning: the acquisition of the Korean three–way stop contrast by native English listeners. This case is particularly interesting in that exposing learners to highly variable stimuli with multiple speaking rates may help them identify which acoustic dimension is more relevant for categorizing the target stop contrast. As discussed below, this hypothesis is based on the fact that the degree of overlap in VOT cues for Korean lenis and aspirated stops varies as a function of speaking rate, thereby reducing the usefulness of VOT for distinguishing these two categories. Thus, HVPT with multiple speaking rates is expected to encourage learners rely on more informative acoustic cues, consequently resulting in more robust training and generalization effects than low–variability training with a single speaking rate.

1.2. Speaking Rate Effects in Korean Stop Contrasts

Unlike English, which has two stop categories (voiced vs. voiceless), Korean has three–way stop contrasts consisting of fortis, lenis, and aspirated categories (e.g., Francis et al., 2008; Harmon et al., 2019). These contrasts are cued by multiple acoustic dimensions, including the VOT and f0 at the onset of the following vowel. While VOT and f0 are used in both English and Korean stop perception, the relative weighting of these cues differs. For the English stop contrasts (e.g., /b/ vs. /p/), VOT is the primary cue, and f0 is secondary (e.g., Francis et al., 2008; Shultz et al., 2012). In contrast, for distinguishing Korean lenis and aspirated stops (e.g., /p/ vs. /p^h/), f0 serves as the primary cue. For older speakers, Korean lenis and aspirated stops are distinguished primarily by differences in VOT, with aspirated stops having relatively longer VOT values than lenis stops (Kang & Guion, 2008). However, recent studies (e.g., Kang, 2014; Lee et al., 2013) report that this distinction has weakened in younger Seoul Korean speakers, leading to a merger of the two categories in terms of VOT. As the VOT difference has reduced, the f0 distinction between lenis and aspirated stops has been enhanced, leading f0 as the primary cue for distinguishing these two categories. Schertz et al. (2015) showed that Korean listeners classified stimuli with lower f0 values more as lenis stops and those with higher f0 values more as aspirate stops, even when VOT values were held constant.[1]

VOT values systematically vary with contextual factors, including speaking rate (Cho & Ladefoged, 1999; Klatt, 1975; Kessinger & Blumstein, 1997). Kessinger & Blumstein (1997) examined the effects of speaking rate on VOT values of stops in English, French, and Thai. Even in the fast–speaking rate condition, category boundaries based on VOT remained relatively stable in these languages, with little overlap between categories (see

1 Fortis stops are primarily distinguished by their shortest VOTs and secondarily by relatively higher f0 compared to lenis stops (Kim, 2004). According to Kang (2014), f0 enhancement was also found for the fortis–lenis contrast (with fortis stops showing relatively higher f0 than lenis stops), but the degree of enhancement was smaller than that for the lenis–aspirated contrast.

Kessinger & Blumstein, 1997, for more details). This stability was interpreted as a mechanism that helps preserve phonetic contrasts among stop categories. However, Korean stop categories exhibit a different pattern. Oh (2009) analyzed the VOTs of Korean fortis, lenis, and aspirated stops produced by 10 native speakers in isolation and at normal and fast speaking rates in a carrier sentence. The results revealed that fortis VOTs remained stable, but VOTs for lenis and aspirated stops overlapped across all conditions. More importantly, the overlap between lenis and aspirated stops increased with faster speaking rates (i.e., isolation: 40.0%, normal: 59.8%, fast: 66.7% for /pin/ and /p^hin/). This overlap has been attributed to the VOT merger, which reduces the motivation to maintain distinct VOT values for these categories across speaking rates.

Based on findings from previous research, a key to successful learning of the Korean stop contrast is understanding the relative importance of VOT and f0 in distinguishing lenis and aspirated stops. Learners must recognize that VOT is a non–diagnostic cue in certain contexts and shift their attention to f0 instead. How, then, can HVPT be modified to promote this learning outcome? One possible approach is to incorporate multiple speaking rates into the training stimuli to increase VOT overlap between lenis and aspirated stops. By exposing learners to greater VOT variability and overlap across speaking rates, we hypothesize that they will learn VOT as an unreliable cue for categorization and instead rely on the more stable cue, f0.

Holt & Lotto (2006) provide theoretical support for this idea. They showed that increasing variability along an uninformative acoustic dimension can restructure listeners' perceptual space, causing attention to shift toward a more informative dimension (see Nosofsky (1987) for the selective attention mechanism resulting from stimulus variability). By increasing the variance of the uninformative acoustic dimension (VOT in the current study), the perceptual distance between training stimuli along that dimension decreased, making it less relevant for categorization decisions. Thus, the current study investigates whether HVPT with increased VOT overlap through multiple speaking rates can help learners identify the more diagnostic acoustic dimension, resulting in higher accuracy in identifying lenis and aspirated stops and in generalizing their learning to novel stimuli.

However, the expected training effects rely on one crucial assumption: f0 is a stable acoustic cue distinguishing lenis and aspirated stops across all speaking rates. Therefore, the present study also conducted an acoustic analysis of the training stimuli to examine whether f0 remains a reliable cue, with a low degree of overlap between lenis and aspirated categories across speaking rates.

The research questions are as follows: (1) Do native English speakers learn the Korean three–way stop contrast, especially lenis and aspirated stops, more effectively with training stimuli that include three different speaking rates (slow, normal, fast) than with a single speaking rate (slow)? (2) Is speaking rate a meaningful dimension for providing phonetic variability in HVPT?

2. Methods

2.1. Participants

Fifteen adult native speakers of American English began the first training session. However, only ten participants (7 female; mean age, 24.5 years; range, 21–38 years) completed the last training session and returned one week later to complete the retention test. Therefore, the current study reports data from these ten participants. All participants were undergraduate or graduate students at a university in the United States and were born and raised in the Midwest. None of the participants reported a history of hearing or speech impairment, nor did any consider themselves bilingual, as indicated by a language background questionnaire. Although some participants had experience with second language learning, none had prior experience with Korean language learning, either formally or informally, qualifying them as naïve learners of Korean at the time of the study. Participants received monetary compensation for their participation.

2.2. High Variability Phonetic Training (HVPT) Stimuli

The training stimuli consisted of 18 bi–syllabic (CVCV) pseudo–words in six minimal triplets. The first syllable of each word combined one of the three Korean bilabial stops, fortis (/p'/), lenis (/p/), or aspirated (/p^h/), with one of the following vowels, /a, e, i, o, u, ʌ/. The second syllable was always /ta/. Because the participants were naïve learners of Korean, the training words were paired with photographs rather than written orthography (e.g., /p'ata/ paired with 'coat', /pata/ with 'zipper', /p^hata/ with 'tree'). All training words in Korean orthography, their IPA transcriptions, and associated pictures can be found in Appendix 1.[2]

Five native Korean female talkers speaking the Seoul dialect (mean age=28.6 years; range=24–34 years) recorded the stimuli. Recordings took place in a sound–attenuated room using a Shure SM–10A microphone at a sampling rate of 44.1 kHz. To create stimuli varying in speaking rates, a metronome was set to three beats per minute (BPM) conditions with a 70 BPM interval: 40 BPM for slow, 110 BPM for normal, and 180 BPM for fast.[3]

The talkers practiced before recording to become comfortable with producing words in sync with the metronome.[4] Each word was produced in isolation multiple times. The best token for each word was selected, resulting in 54 stimuli per talker (18 words×3 rates). Stimuli from four talkers (Talkers 1, 2, 3, & 4) were used for the training sessions, and those from the remaining talker (Talker 5) were used for the new talker generalization test.

2.3. High Variability Phonetic Training (HVPT) Tasks and Procedure

The HVPT spanned three days and consisted of three training

2 All appendices for this manuscript are available on the Open Science Framework at [https://osf.io/vydte/].

3 This recording method was inspired by Miller & Baer (1983), who examined whether the transition durations of /b/ and /w/ change as a function of speaking rate. In their study, a metronome ranging from 40 to 192 BPM was used to induce slow and fast speech rates.

4 An anonymous reviewer noted that metronome pacing may also affect other prosodic properties (e.g., segmental durations, coarticulation, F0 contour). We acknowledge this as a potential limitation, although our focus here is on VOT and f0.

sessions. Once participants began their first training session, they were required to continue through the final day without interruption. All participants completed the training without any interim breaks.

On the first day of participation, participants were randomly assigned to one of two groups: the Multi–group or the Single group. Although all participants were naïve learners of Korean, a pre–oddball test was administered before the HVPT to ensure that they had comparable auditory sensitivity to the target Korean stop contrast. The stimuli were taken from one of the training talkers (Talker 2) produced at slow–rate. On each trial, participants first heard three stimuli in a row (e.g., /pa/–/pa/–/p'a/) and then selected the token that was different from the other two. The percentage of correct responses was calculated out of 108 trials [6 vowels×3 presentation orders (ABB, AAB, ABA)×6 AB pairs (/p–p^{h}/, /p–p'/, /p^{h}–p'/, /p^{h}–p/, /p'–p/, /p'–p^{h}/)]. An independent *t*–test revealed no significant difference between the Multi–group (*M*=93.7%, *SD*=4.4%) and the Single group (*M*=89.8%, *SD*=5.8%), *p*=.27. This result suggests that both groups had similar sensitivity to the target stop contrast before beginning the training.

Each training session began with a daily familiarization phase, during which all 18 target words were presented auditorily along with their corresponding photographs. Only slow–rate stimuli from one training talker were used in this phase (18 words×1 rates×2 repetitions=36 trials). The main training session was followed in the form of a three–alternative forced–choice (3AFC) identification (ID) task. On each trial, an auditory stimulus was played over headphones, accompanied by three photographs displayed on the computer screen. It should be noted that the photographs presented together in each trial shared the same vowel in the first syllable (e.g., /p'ida/, /pida/, and /p^{h}ita/). Participants were asked to click on the photograph corresponding to the word they heard. Trial–by–trial feedback was provided; if a response was incorrect, the auditory stimulus with the correct picture was presented again.

The Multi–group was trained with stimuli produced by four talkers across all speaking rates (slow, normal, and fast) (18 words×3 rates×4 talkers=216 trials). The Single group was trained with stimuli from the same four talkers but only at the slow rate. To equate the total number of training trials, stimuli were repeated three times for the Single group (18 words×1 rate×4 talkers×3 repetitions=216 trials).

Training sessions were divided into four blocks, with each block containing stimuli from only one talker, rather than mixing all voices, to enhance the effectiveness of HVPT (Perrachione et al., 2011). Stimuli within each block were presented in a randomized order. Each training session concluded with an everyday 3AFC ID test (henceforth, everyday ID test) to track participants' learning progress throughout the training. This test followed the same procedure as the training phase, except no feedback was given. Only the slow–rate stimuli from Talker 1, with two repetitions per word (18 words×1 slow–rate×1 talker×2 repetitions=36 trials), were used. Each session (training and the everyday ID test) lasted about one hour.

Participants' generalization of their learning to a new voice was assessed on the final day of training using the new talker generalization task. This test used slow–rate stimuli produced by Talker 5, who had not been used in the training phase (18 words×1 slow–rate×1 novel talker×2 repetitions=36 trials). One week after the last training session, participants returned for the retention test. This test repeated both the everyday ID test and a new talker generalization test to evaluate retention of learning.

2.4. High Variability Phonetic Training (HVPT) Stimuli Acoustic Analysis

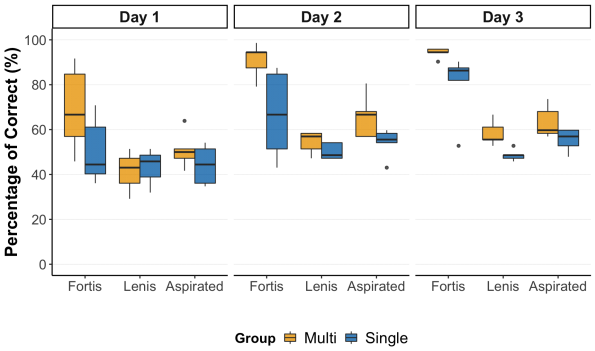
As noted earlier, it is essential to confirm whether the training stimuli remained f0 relatively distinct between lenis and aspirated stops across speaking rates with minimal overlap. Therefore, we conducted an acoustic analysis with training stimuli recorded by four training talkers. Each stimulus was analyzed using Praat (Boersma, 2001). VOT was measured from the beginning of the initial bilabial consonant burst to the first zero crossing in the waveform following the onset of periodicity in the following vowel (Schertz et al., 2015). f0 was measured at 5 ms after the onset of periodicity in the vowel using the "To Pitch..." function in Praat (Schertz et al., 2015). Following Oh (2009), we examined how the overlap of VOT and f0 values for each Korean stop category varied by speaking rates (slow, normal, fast).

3. Results

3.1. High Variability Phonetic Training (HVPT) Performance

Figure 1 shows box plots of identification accuracies on each day of training for the Multi and Single groups. Each panel separately presents the accuracy for each Korean target stop (fortis, lenis, aspirated). Visual inspection and mean values suggest that both groups show a trend of identifying the Korean fortis stops better than the other two types of stops as they received more training (Multi: 69.2%, 90.8%, 94.2%; Single: 50.6%, 66.7%, 79.8%). However, the Multi–group appeared to learn to correctly identify fortis stops faster than the Single group, as indicated by the reduced individual variability among the Multi–group participants. If we compare the interquartile ranges of box plots for fortis stops, the ranges in the Multi–group decreased considerably from Day 1 to Day 2 compared to the Single group. Although the Single group also eventually learned to identify fortis stops by Day 3, Figure 1 suggests that HVPT with multiple speaking rates facilitates faster learning of Korean fortis stops compared to training with a single speaking rate.

Compared to fortis stops, both groups showed overall lower mean identification accuracies and slower improvement trajectories for Korean lenis (Multi: 41.1%, 54.4%, 56.9%; Single: 43.3%, 50.3%, 48.6%) and aspirated stops (Multi: 50.8%, 65.8%, 63.6%; Single: 44.2%, 54.2%, 55.2%). Nevertheless, beginning on Day 2, the Multi–group and the Single group began to diverge in their learning patterns. The Multi–group identified Korean lenis and aspirated stops more accurately than the Single group, although these group differences were less pronounced than for fortis stops.



Boxplots: The boxed region indicates the interquartile range; whiskers extend to extreme values; the solid bar indicates the median; points outside of whiskers indicate outliers.

Figure 1. Identification accuracy of each day of the training session for the Multi and the Single groups.

To assess the overall group differences across training days, we conducted mixed–effects logistic regression analyses using participants' identification response data (Jaeger, 2008). We used the *glmer*() function from the *lme4* package (version 1.1–27) in *R* (R Core Team, 2024). The binary outcome variable represented correct (1) or incorrect (0) responses. Fixed effects included Group (Single vs. Multi), Training Day (Day 1, Day 2, Day 3), and their interactions. Group was centered (–0.5 and 0.5) in order that the main effects were evaluated as the average effects over all levels of Group. Training Days were dummy–coded with Day 1 as the reference level (Day 1 VERSUS Day 2, Day 1 VERSUS Day 3). In this manuscript, we focused on reporting only the results that are discussed—primarily significant or marginally significant main effects and interactions. The full model outcomes are provided in Appendix 2 (top).

Table 1 (top) summarizes the overall results of training sessions. The significant main effect of Group indicates an overall higher performance by the Multi–group averaged across training days. Furthermore, the significant main effects of Day 1 VERSUS Day 2 and Day 1 VERSUS Day 3 demonstrate participants' improvements over time as they receive more training sessions. A significant interaction between Group and Day 1 VERSUS Day 2 suggests that the magnitude of improvement on Day 2 compared to Day 1 was greater for the Multi–group than the Single group. These findings provide preliminary evidence that exposure to varied speaking rates may promote more effective perceptual learning of L2 contrasts.

As discussed earlier, the Korean lenis and aspirated stop distinction relies more on f0 than VOT for identification, unlike the English stop contrast, where VOT serves as the primary cue. Therefore, participants were expected to experience more difficulty identifying lenis and aspirated stops. In addition, both Single and Multi groups reached a celling effect in identifying fortis stops by the last day of training. To examine which training type is more effective in facilitating the identification of Korean lenis and aspirated contrast, we fit another mixed–effects logistic regression model with Contrast as a fixed effect (centered as –0.5=Lenis, 0.5=Aspirated). The full model outcomes are provided in Appendix 2 (bottom). In Table 1 (bottom), the significant main effect of Contrast suggests that aspirated stops were identified more accurately than lenis stops. A possible reason for this finding is discussed in Section 4. The model also found a significant interaction between Group and Contrast, which shows that the Multi –group identified the aspirated stops more accurately than the lenis stops compared to the Single group.

Table 1. The output of logistic regression models for the overall training session results (top) and results for lenis and aspirated tops (bottom)

Fixed Effects	Estimate	*SE*	z	*p*–value
(Intercept)	0.42	0.08	5.25	<0.001
Group	0.46	0.05	8.77	<0.001
Day1 VERSUS Day2	0.58	0.08	7.34	<0.001
Day1 VERSUS Day3	0.70	0.09	7.66	<0.001
Group×Day1 VERSUS Day2	0.27	0.13	2.17	0.030
Fixed Effects	Estimate	SE	z	*p*–value
(Intercept)	0.10	0.06	1.80	0.072
Contrast	0.24	0.06	3.95	<0.001
Day1 VERSUS Day2	0.46	0.08	6.08	<0.001
Day1 VERSUS Day3	0.46	0.08	6.07	<0.001
Group×Contrast	0.36	0.12	2.95	0.003

3.2. Everyday ID Tests and New Talker Generalization Test

Figure 2 shows the results of the everyday ID tests, administered after each training session, as well as the new talker generalization test. The results align with the training session patterns described in Section 3.1. The Multi–group outperformed the Single group on everyday ID tests. This pattern reflects the same group advantage observed during the training sessions, particularly for the more difficult lenis and aspirated contrasts. While both groups successfully identified Korean fortis stops, performances for lenis and aspirated stops differed. The Multi–group showed consistent improvement across the subsequent everyday ID tests compared to Day 1 (lenis: 56.6%, 60%, 75%; aspirated: 70%, 83%, 78%), whereas the Single group's performance remained relatively similar to their Day 1 everyday ID test (lenis: 46.6%, 48.3%, 61.6%; aspirated: 51.6%, 51.6%, 48.3%). For the generalization test, the Multi–group also showed better generalization to stimuli produced by a novel talker (fortis: 95%, lenis: 73.3%, aspirated: 71.6%) than the Single group (fortis: 80%, lenis: 58.3%, aspirated: 40%).

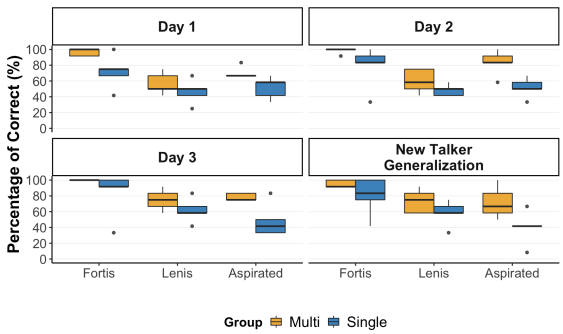


Figure 2. Identification accuracy of everyday identification (ID) tests and new talker generalization test for Single and Multi groups.

We built mixed–effects logistic regression models with Group (centered as –0.5=Single, 0.5=Multi), Training Day (reference level: Day 1), and their interactions as fixed effects. One model was fitted to the overall everyday ID test results, and an additional model was built to examine performance on lenis and aspirated stops with the Contrast variable (see Appendix 3 for full models). As shown in Table 2, both models found a significant main effect of Group. Notably, the main effect of Group in the model with the Contrast

variable—which was not observed in the training session analysis in Section 3.1—shows that the Multi–group identified lenis and aspirated stops more accurately than the Single group. This pattern may be because the everyday ID tests did not provide trial–by–trial feedback as the training sessions did. When feedback was absent, the Multi–group's better performance appeared to be more evident. The model with the Contrast variable also found a significant interaction between Group and Contrast, indicating that the greater accuracy in identifying aspirated stops compared to lenis stops was larger in the Multi–group than in the Single group.

No significant interactions between Group and Training Day were observed, possibly due to the small sample size (N=10) in the current study. To further explore potential group differences that may not have been fully captured by the interaction term, we conducted post hoc tests using the *emmeans*() function on the response scale. Results showed that the Multi–group outperformed the Single group on the Day 1 ID test (p=.003). This advantage persisted on Day 2 (p<.001) and Day 3 (p<.001), and the new talker generalization test (p<.001). However, it should be emphasized that given the small sample size, these post hoc results should be considered exploratory and interpreted with caution.

Table 2. The output of logistic regression models for the overall everyday identification (ID) and new talker generalization test results (top) and results for lenis and aspirated tops (bottom)

Fixed Effects	Estimate	*SE*	*z*	*p*–value
(Intercept)	0.91	0.10	9.43	<0.001
Group	0.99	0.19	5.12	<0.001
Day1 VERSUS Day3	0.48	0.17	2.77	0.006
Fixed Effects	Estimate	*SE*	*z*	*p*–value
(Intercept)	0.93	0.10	9.48	<0.001
Group	1.02	0.20	5.20	<0.001
Day1 VERSUS Day2	0.29	0.17	1.7	0.097
Day1 VERSUS Day3	0.47	0.17	2.73	0.006
Group×Contrast	0.88	0.31	2.87	0.004

3.3. Retention Test

Figure 3 shows the retention test result, which replicated the everyday ID and new talker generalization tests one week after training. We examined how much participants retained their learning and compared their retention performances to Day 3 (the last day of training). Overall, the results mirrored those observed during training sessions and everyday ID tests. The Multi–group again performed better than the Single group in both sessions of retention tests, one with the old talker and the other with a new talker.

Two separate logistic regression analyses results for the retention tests are presented in Appendix 4. Fixed effects included Group and Test [Day 3 (reference–level) vs. retention]. Both models revealed a significant main effect of Group, with the Multi–group showing overall higher scores on the tests with both old and new talkers. There was a marginal Group and Test interaction in the model with the old talker (p=.084), suggesting that the Multi–group showed a larger drop in accuracy from Day 3 to the retention test than the Single group, although the Multi–group still scored higher overall. One thing to note here is that the Single group performed better on the retention test compared to the last day of training when tested with the old talker. Although this difference was not statistically significant, the trend may reflect memory consolidation of learning (Tamminen et al., 2012).

Figure 3. Identification accuracy of retention test for old talker [same as everyday identification (ID) test] and new talker (same as new talker generalization test) for Single and Multi groups.

3.4. Acoustic Analysis of High Variability Phonetic Training (HVPT) Stimuli

This section reports the results of an acoustic analysis of the HVPT stimuli produced by four training talkers. Recall that the goal of this analysis was to examine how the distributions of VOT and f0 values for fortis, lenis, and aspirated stops varied across different speaking rates, and whether the degree of overlap between these distributions was affected by speech rates (Oh, 2009).

Figure 4 presents the VOT distributions of training stimuli. For lenis and aspirated stops (left), there is a notable degree of VOT overlap across all speaking rates. As the speaking rate increased, VOT values shortened, the overlap became more pronounced (overlap%: slow 52%, normal 54%, fast 78%), and the category boundary between lenis and aspirated stops became more ambiguous. In contrast, the VOT values for fortis stops remained relatively stable across speaking rates. These patterns replicate the findings of Oh (2009), demonstrating that speaking rate significantly affects category distinctions on the VOT dimension, particularly losing its role in contrasting lenis and aspirated stops.

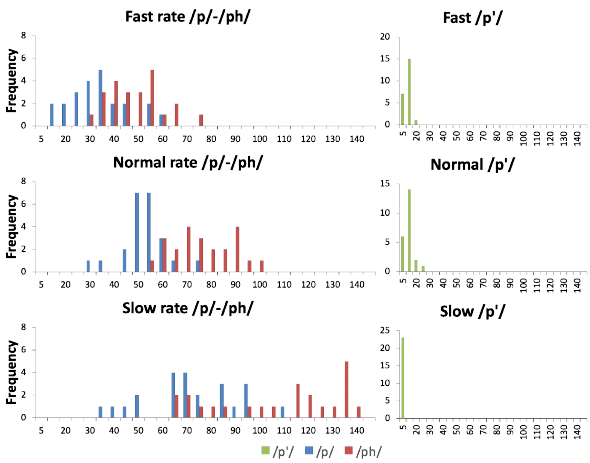


Figure 4. Voice Onset Time (VOT) distributions (in ms) of fortis (right), lenis, and aspirated (left) stops across speaking rates: fast (top), normal (middle), slow (bottom).

When L2 learners are trained with stimuli in which one acoustic

dimension (such as VOT) exhibits considerable overlap between categories, they are likely to rely on another acoustic dimension that more reliably signals the target contrast. In the case of the lenis and aspirated distinction in Korean, f0 is known to be a more informative cue. Therefore, it is important that the training stimuli provide reliable f0 distinctions between these categories. To assess the reliability of f0 as a cue signaling lenis and aspirated distinction in our training stimuli, we examined the range and degree of overlap in f0 between these stops across speech rates for each training talker.

Figure 5 illustrates the f0 distributions of stops across speech rates for each talker (see also Appendix 5). The f0 values for the lenis (blue) and aspirated (dark red) categories were consistently well–separated, with relatively minimal overlap, even in the fast–rate condition. The lenis and aspirated contrast was maintained clearly regardless of speaking rates, supporting the idea that f0 served as a stable and reliable cue across speaking rate conditions in this study. Overall, the results of the acoustic analysis suggest that during HVPT involving multiple speaking rates, learners were exposed to stimuli in which VOTs were highly overlapping for lenis and aspirated stops, while f0 values remained distinct. As a result, the role of f0 was reinforced during training in correctly identifying Korean lenis and aspirated stops.

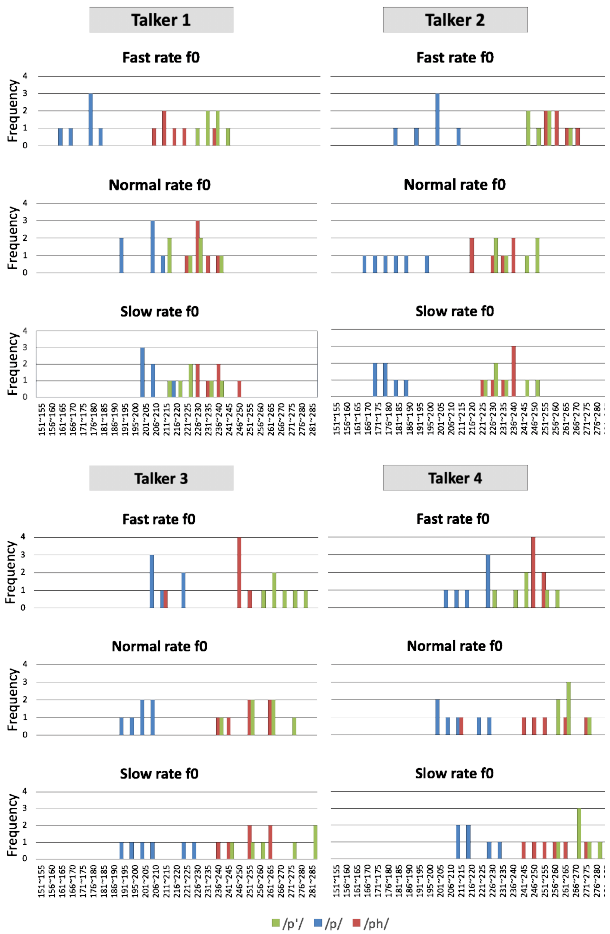


Figure 5. f0 distributions (in Hz) of fortis, lenis, and aspirated stops across speaking rates for each training talker: fast, normal, slow.

4. Discussion

The present study investigated the effectiveness of a particular type of HVPT in facilitating the acquisition of the Korean three–way stop contrast by native English speaking naïve learners. Specifically, we examined whether increased variability in training stimuli, implemented through multiple speaking rates (slow, normal, fast), could promote more effective learning of L2 phonological contrasts. Training effects were assessed through a series of tests, including everyday ID tests, a new talker generalization test, and a retention test. Overall, our findings support the hypothesis that HVPT with increased variability in speaking rates leads to more robust perceptual learning of L2 contrasts. It is important to note that both groups in our study received HVPT training with stimuli produced by multiple talkers. Although both groups improved over time, the Multi–group showed greater gains during training, performed better on tests, and exhibited faster learning compared to the Single group.

These findings align with previous research demonstrating that HVPT is an effective paradigm for training non–native speech contrasts (e.g., Bradlow et al., 1999; Lively et al., 1993; Logan et al., 1991). However, the current study extends prior work by identifying speaking rate as a meaningful dimension of training stimulus variability (Research Question 1), in addition to more commonly studied factors such as talker and phonetic context variability. We acknowledge that some of the statistical effects were marginal, but the overall trend suggests that participants exposed to multiple speaking rates demonstrated greater training advantages than those exposed to a single rate. The potential benefits of HVPT with multiple speaking rates were reflected not only in training performance but also in the tendency for the Multi–group to perform better in generalization to an unfamiliar talker and in retention of training effects one week later compared to the Single group (Research Question 2). While the Multi–group's training advantage was supported by both descriptive and inferential analyses, not all interactions, particularly those involving Group and Training Days variables, reached statistical significance. One likely reason is the small sample size, which may have limited statistical power. Larger –scale studies are needed to more conclusively evaluate the robustness of speaking rate variability as a mechanism for enhancing HVPT effectiveness.

It is worth noting that the Single group was trained exclusively with slow–rate stimuli, which were repeated three times and used in all tests for both groups. Given this, one might expect the Single group to outperform the Multi–group on the tests due to their greater exposure to test–matching stimuli. However, the Multi–group still outperformed the Single group despite having less exposure to slow–rate input. This suggests that learning outcomes cannot be enhanced merely by increasing the amount of exposure. Rather, incorporating stimulus variability should be considered a key component in designing L2 training paradigms for more effective learning.

Although previous studies have yielded mixed results regarding the potential benefits of speaking rate variability in HVPT (Hirata et al., 2007; Sonu et al., 2013), our findings align with those of Hirata et al. (2007), who found marginal benefits of mixed speaking rate training on the perception of Japanese vowel length contrasts. Importantly, the current study addressed some limitations raised in Hirata et al. (2007) by

incorporating three speaking rates with equal temporal intervals (70 BPM). However, the two studies differed in their goals for employing high stimulus variability with multiple speaking rates. In Hirata et al. (2007), the primary focus was on training *rate normalization*—an essential perceptual strategy for identifying relative durational differences between Japanese short and long vowels by referencing durations of surrounding segments. Their use of slow and fast speaking rates in training was intended to help learners recognize how durational cues signaling the target contrast systematically vary with speaking rates, thereby encouraging them to avoid over–reliance on absolute durational differences. In contrast, the current study used multiple speaking rates in training not to demonstrate how an important acoustic cue systematically varies with speaking rate, but to increase *ambiguity* in one acoustic dimension (VOT) to which native English learners are biased due to their L1 experience.

Since L2 learners often struggle to determine which acoustic dimensions are most relevant for identifying non–native contrasts and to redirect their attention accordingly (Idemaru & Holt, 2011; Schertz et al., 2015), we incorporated multiple speaking rates in creating training stimuli to reduce the reliability of VOT and thereby increase the informativeness of f0. Our findings, showing higher identification accuracy of lenis and aspirated stops by the Multi–group, provide preliminary support for the prediction that participants may have shifted their attention to f0. Nonetheless, this interpretation remains speculative. We did not directly measure learners' changes in their cue–weighting strategies before and after the training, nor did we examine perceptual attention shifts from VOT to f0 over the course of training. Future research should incorporate cue–weighting tasks to examine whether training with multiple speaking rates encourages greater reliance on f0 over VOT. Such evidence would strengthen the claim that high stimulus variability can guide perceptual reallocation to more informative acoustic dimensions.

With regard to learning patterns across phonation types, fortis stops were more easily acquired than lenis and aspirated stops in both groups. This finding is consistent with existing L2 speech perception models (e.g., PAM–L2, Best & Tyler, 2007) and cross–language mapping studies (Lee & Park, 2024; Schmidt, 2007), which emphasize the role of phonetic similarities and dissimilarities between L1 and L2 categories in determining the relative difficulty of learning L2 contrasts. For English listeners, both Korean lenis and aspirated stops are often mapped onto the single English voiceless stop category (i.e., Single Category assimilation), whereas Korean fortis stops are more likely to be mapped onto the homorganic English voiced stop category (Schmidt, 2007). The difficulty in acquiring the lenis and aspirated distinction, therefore, likely stems from the Single Category assimilation, as well as from the reversed relationship between the primary and secondary acoustic dimensions in the two languages. One notable learning pattern is that Korean aspirated stops were identified more accurately than lenis stops. This result might be contributed to that Korean aspirated stops are signaled by relatively higher f0 than lenis stops, which aligns with the f0 cue use for English voiceless stops. It is speculated that the similar way of using f0 cues to signal Korean aspirated and English voiceless stops might help learners identify aspirated stops better than lenis stops, particularly among for the naïve learners of Korean. Future longitudinal research on Korean three–way stop learning trajectories is needed to further clarify this pattern.

Individual differences in perceptual learning further complicate the effectiveness of HVPT. Variability in learning success has been widely documented and attributed to a range of factors, including cognitive abilities (Darcy et al., 2015), native language processing (Lengeris, 2009), and cue–weighting strategies (Lee & Park, 2024). Although the present study demonstrated group–level advantages for multiple speaking rate training, not all learners may benefit equally from this training paradigm. For some, increased variability may impose processing difficulty or reduce engagement (e.g., Giannakopoulou et al., 2017; Perrachione et al., 2011). Future research should explore how individual differences, such as perceptual gradiency, working memory, and cognitive control, interact with HVPT involving speaking rate variability. Moreover, tailoring HVPT to better accommodate individual learner traits will be a valuable direction for future research.

References

Best, C. T., & Tyler, M. D. (2007). Nonnative and second–language speech perception: Commonalities and complementarities. In M. J. Munro, & O. S. Bohn (Eds.), *Language experience in second language speech learning: In honor of James Emil Flege* (pp. 13–34). Amsterdam: John Benjamins.

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International*, *5*(9), 341–345.

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

Cho, T., & Ladefoged, P. (1999). Variation and universals in VOT: Evidence from 18 languages. *Journal of Phonetics*, *27*(2), 207–229.

Darcy, I., Park, H., & Yang, C. L. (2015). Individual differences in L2 acquisition of English phonology: The relation between cognitive abilities and phonological processing. *Learning and Individual Differences*, *40*, 63–72.

Francis, A. L., Kaganovich, N., & Driscoll–Huber, C. (2008). Cue –specific effects of categorization training on the relative weighting of acoustic cues to consonant voicing in English. *The Journal of the Acoustical Society of America*, *124*(2), 1234–1251.

Giannakopoulou, A., Brown, H., Clayards, M., & Wonnacott, E. (2017). High or low? Comparing high and low–variability phonetic training in adult and child second language learners. *PeerJ*, *5*, e3209.

Harmon, Z., Idemaru, K., & Kapatsinski, V. (2019). Learning mechanisms in cue reweighting. *Cognition*, *189*, 76–88.

Hirata, Y., Whitehurst, E., & Cullings, E. (2007). Training native English speakers to identify Japanese vowel length contrast with sentences at varied speaking rates. *The Journal of the Acoustical Society of America*, *121*(6), 3837–3845.

Holt, L. L., & Lotto, A. J. (2006). Cue weighting in auditory categorization: Implications for first and second language

acquisition. *The Journal of the Acoustical Society of America*, *119*(5), 3059–3071.
Idemaru, K., & Holt, L. L. (2011). Word recognition reflects dimension–based statistical learning. *Journal of Experimental Psychology: Human Perception and Performance*, *37*(6), 1939–1956.
Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*(4), 434–446.
Kang, K. H., & Guion, S. G. (2008). Clear speech production of Korean stops: Changing phonetic targets and enhancement strategies. *The Journal of the Acoustical Society of America*, *124*(6), 3909–3917.
Kang, Y. (2014). Voice Onset Time merger and development of tonal contrast in Seoul Korean stops: A corpus study. *Journal of Phonetics*, *45*, 76-90.
Kessinger, R. H., & Blumstein, S. E. (1997). Effects of speaking rate on voice–onset time in Thai, French, and English. *Journal of Phonetics*, *25*(2), 143–168.
Kim, M. (2004, October). Correlation between VOT and F0 in the perception of Korean stops and affricates. *Proceedings of the 8th International Conference on Spoken Language Processing (INTERSPEECH–2004)* (pp. 49–52). Jeju, Korea.
Klatt, D. H. (1975). Voice onset time, frication, and aspiration in word–initial consonant clusters. *Journal of Speech and Hearing Research*, *18*(4), 686–706.
Lee, H., Politzer–Ahles, S., & Jongman, A. (2013). Speakers of tonal and non–tonal Korean dialects use different cue weightings in the perception of the three–way laryngeal stop contrast. *Journal of Phonetics*, *41*(2), 117–132.
Lee, J., & Park, H. (2024). Acoustic cue sensitivity in the perception of native category and their relation to nonnative phonological contrast learning. *Journal of Phonetics*, *104*, 101327.
Lengeris, A. (2009). *Individual differences in second–language vowel learning* (Doctoral Dissertation). University College London, London, UK.
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. *The Journal of the Acoustical Society of America*, *94*(3), 1242–1255.
Logan, J. S., Lively, S. E., & Pisoni, D. B. (1991). Training Japanese listeners to identify English /r/ and /l/: A first report. *The Journal of the Acoustical Society of America*, *89*(2), 874–886.
Miller, J. L., & Baer, T. (1983). Some effects of speaking rate on the production of /b/ and /w/. *The Journal of the Acoustical Society of America*, *73*(5), 1751–1755.
Nosofsky, R. M. (1987). Attention and learning processes in the identification and categorization of integral stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *13*(1), 87–108.
Oh, E. J. (2009). Voice onset time of Korean stops as a function of speaking rate. *Phonetics and Speech Sciences*, *1*(3), 39–48.
Perrachione, T. K., Lee, J., Ha, L. Y. Y., & Wong, P. C. M. (2011). Learning a novel phonological contrast depends on interactions between individual differences and training paradigm design. *The Journal of the Acoustical Society of America*, *130*(1), 461–472.
R Core Team. (2024). R: A language and environment for statistical computing (version 4.3.3) [Computer software]. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from https://www.R–project.org/
Schertz, J., Cho, T., Lotto, A., & Warner, N. (2015). Individual differences in phonetic cue use in production and perception of a non–native sound contrast. *Journal of Phonetics*, *52*, 183–204.
Schmidt, A. M. (2007). Cross–language consonant identification: English and Korean. In O. S. Bohn, & M. J. Munro (Eds.), *Language experience in second language speech learning* (pp. 185–200). Amsterdam, Netherlands: John Benjamins.
Shultz, A. A., Francis, A. L., & Llanos, F. (2012). Differential cue weighting in perception and production of consonant voicing. The *Journal of the Acoustical Society of America, 132*(2), EL95–EL101.
Sonu, M., Kato, H., Tajima, K., Akahane–Yamada, R., & Sagisaka, Y. (2013). Non–native perception and learning of the phonemic length contrast in spoken Japanese: Training Korean listeners using words with geminate and singleton phonemes. *Journal of East Asian Linguistics, 22*(4), 373–398.
Tamminen, J., Davis, M. H., Merkx, M., & Rastle, K. (2012). The role of memory consolidation in generalisation of new linguistic information. *Cognition, 125*(1), 107–112.

● **Jieun Lee,** Corresponding author
Visiting Assistant Professor, Dept. of Linguistics, University of Kansas
1541 Lilac Lane, Lawrence, KS 66045, USA
Tel: +1-785-864-5226
Email: jieunlee@ku.edu
Areas of interest: Phonetics, L2 acquisition

● **Hanyong Park**
Associate Professor, Dept. of Linguistics, University of Wisconsin-Milwaukee
Johnston Hall 123, 2522 E Hartford Ave, Milwaukee, WI 53211, USA
Tel: +1-414-251-8789
Email: park27@uwm.edu
Areas of interest: Phonetics, L2 acquisition, Sociophonetics

Appendix

The appendix is available at https://osf.io/vydte/.