Perceptual Training Enhances the Use of Vowel Quality Cues to Lexical Stress: The Benefits of Intonational Variability

Annie Tremblay,[1] Hyoju Kim, [2] Sahyang Kim, [3] & Taehong Cho[4]

1. The University of Texas at El Paso, 2. University of Kansas, 3. Hongik University, 4. Hanyang University
actremblay@utep.edu, kimhj@ku.edu, sahyang@hongik.ac.kr, tcho@hanyang.ac.kr

ABSTRACT

This study investigates whether high-variability phonetic training, also known as multi-talker phonetic training, enhances Seoul Korean listeners' weightings of acoustic cues to English lexical stress and does so more than single-talker perceptual training. Seoul Korean listeners at an intermediate proficiency in English completed a cue-weighting stress perception task (pre-test), eight perceptual training sessions (over eight consecutive days) in which they heard noun-verb stress minimal pairs produced by one or four talkers and identified the word they heard, and the same cue-weighting stress perception task (post-test). In both training conditions, the stimuli varied in their intonational realizations. The results showed that both training types similarly enhanced Korean listeners' use of vowel quality cues to English lexical stress, and training increased listeners' use of pitch cues in the absence of vowel quality cues. The comparable effects of training type are attributed to the intonational variability in the training stimuli.

Keywords: Stress perception, cue weighting, perceptual training, Korean, English

1. INTRODUCTION

High-Variability Phonetic Training (HVPT), also known as multi-talker phonetic training, has been shown to be highly successful for improving listeners' discrimination and identification of difficult second-language (L2) sound contrasts, more so than single-talker phonetic training (STT) [1, 2]. HVPT has been shown to enhance listeners' perception of L2 phonetic categories [3], L2 syllable structure [4], and L2 lexical tones [5]. Phonetically variable speech has been deemed beneficial to L2 speech learning because it enables listeners to weight multiple dimensions of linguistic contrasts relative to the phonetic context in which they are heard, thus aiding listeners' development of robust L2 perceptual representations. What is less clear from this research, however, is whether the benefits of HVPT extend to the weighting of acoustic cues to lexical stress for listeners whose first language (L1) does not have lexical stress. The present study seeks to answer this question with Seoul Korean L2 learners of English.

English has lexical stress. For example, in English, words such as *DEsert* and *deSSERT* (with capitalized letters representing the stressed syllables) have different stress patterns. The stress contrast in noun-verb pairs is reflected in the alternation of full-reduced vowels. Since full and reduced vowels are realized differently in the spatial dimension, the primary acoustic correlate of lexical stress is often considered to be vowel quality [6]. Importantly, the full-reduced forms are also distinguished in the temporal dimension, such that a difference in vowel quality is accompanied by an intrinsic difference in duration, with full vowels being longer than reduced ones. Since a difference in duration does not necessarily induce a difference in vowel quality (unlike the reverse), duration is considered a secondary phonetic correlate of lexical stress [7]. Pitch is also known to be an important correlate of lexical stress [8], but pitch is realized differently as a function of phrase-level pitch accent types (L*, H*, L+H*, L*+H) [9,10], such that there is no one-to-one relationship between pitch and lexical stress, making the pitch change a less consistent correlate. In contrast, Seoul Korean does not have lexical stress. Prominence is realized intonationally by phrasal edge tones, with the Accentual Phrase (AP) having the underlying LHLH or HHLH tonal pattern (where L = low and H = high). The first tone of the AP varies as a function of the phrase-initial segment (H for fortis and aspirated segments, and L for all other segments) [11]. Consequently, pitch is an intonationally driven correlate of this segmental (also lexical) contrast.

Previous research has shown that Korean L2 learners of English (L1 dialect(s) unspecified) have more difficulty recalling sequences of English nonwords differing in stress, where the stress contrast is realized primarily with suprasegmental cues (e.g., ['mipa] vs. [mi'pa]), than sequences of English nonwords differing in a segment (e.g., ['kupi] vs. ['kuti]) [11]. However, research findings differ as to whether Korean L2 learners of English can use vowel quality cues to lexical stress in spoken word recognition. Lin et al. [12] report that Korean L2 learners of English are not more accurate at rejecting nonwords whose incorrect stress placement is signaled by vowel quality cues (e.g., **HOrizon* ['hɑɪaɪzən]) compared to incorrectly stressed nonwords without vowel quality changes

(e.g., **Enough* ['ɪnʌf]). By contrast, Connell et al. [13] found that Korean L2 learners of English show higher target-over-competitor fixation proportions when the first syllable of the target and competitor differ segmentally and suprasegmentally (e.g., *PArrot* vs. *paRADE*) than when they do not (e.g., *PArrot* vs. *PArish*), an effect not found for target and competitor words whose first syllable differ only supra-segmentally (e.g., *SURface* vs. *surPRISE* or *SURplus*). To explain this, the authors proposed that Korean L2 learners of English may have assimilated full and reduced English vowels to different Korean vowels and use vowel quality differences to distinguish target from competitor words. In other words, Korean listeners may transfer the use of spectral cues from the perception of vowels in Korean to the perception of lexical stress in English.

One important remaining issue, however, is whether HVPT can enhance the weighting of acoustic cues to English lexical stress in listeners whose L1 does not have lexical stress, and result in more target-like cue weighting than STT. Given Seoul Korean listeners' sensitivity to pitch as a prosodic cue that signals a segmental contrast (lenis vs. fortis and aspirated segments in phrase-initial position), we might expect Korean listeners to rely on pitch cues to English lexical stress. What remains to be seen is whether HVPT can help them rely less on pitch and more on vowel quality when perceiving English lexical stress. The present study will elucidate whether this is the case, comparing the efficiency of HVPT and STT for enhancing cue-weighting. Since this is the first study that seeks to answer this question, it is unclear whether perceptual training should target a specific cue distribution or whether a distribution of cues that mimics spoken English (as established from corpus studies) would be sufficient to enhance learning. We opted for the latter as a starting point into this investigation. The cue-weighting task that served as pre- and post-test in this study is the one that Tremblay et al. [14] used to test Dutch listeners' perception of English lexical stress.

2. METHOD

2.1. Participants

The participants were 54 Seoul Korean L2 learners of English (mean age: 24, 32 females) (for a comparison with native English listeners, see [14]). The Korean listeners were tested at a Korean university in Seoul. Among them, 27 were randomly assigned to HVPT and 27 to STT. All participants completed a detailed language background questionnaire and the Lexical Test for Advanced Learners of English (LexTALE) [15] to measure lexical proficiency in English (mean: 68.1, SD: 8.4). The two L2 subgroups did not differ significantly in their age, age of first exposure to English, years of English education, English proficiency self-ratings, English accent self-ratings, or LexTALE score. No participant reported a history of speech, language, or hearing impairments.

2.2. Materials

2.2.1. Training stimuli

The words that served as training stimuli were noun-verb minimal pairs that differed in lexical stress. The lexical items were 28 English noun-verb pairs for which at least one of the two vowels was reduced when unstressed (e.g., *REcall* vs. *reCALL*), and 8 English noun-verb minimal pairs without vowel reduction (e.g., *PERmit* vs. *perMIT*). This distribution of words with and without vowel reduction cues (78% vs. 22%) was based on Cutler and Carter [16]'s corpus study. The token frequency of the words as a noun and that of a verb was controlled based on the Corpus of Contemporary American English [17].

The auditory stimuli for HVPT were recorded by four native speakers of American English (two male and two female). The auditory stimuli for STT were recorded by a female speaker who also recorded items for HVPT. The 36 word pairs were elicited and recorded with three different intonations (H*L–, L*H–, and flat intonational contour). The target words with H*L– were elicited in the declarative sentence *Mary said ____ before*. The words with L*H– were elicited in the interrogative sentence *Mary said ____ before?* The words with a flat intonational contour were elicited in the carrier phrase *MARY said ____ before*, where *MARY* had a contrastive pitch accent and where the target word was deaccented. The intensity of the words was normalized to 70 dB.

2.2.2. Pre-/post-test stimuli

The word pair that was used for the cue-weighting stress perception task that served as pre- and post-test was *DEsert-deSSERT.* It was recorded with a H* pitch accent by a female native speaker of American English who did not record the training stimuli; thus, any learning from the training is evidence for generalization to a new talker. The word pair was elicited in the carrier sentence *Click on* ____. One token of *DEsert* served as the base token and was manipulated to have seven steps for each acoustic dimension (vowel quality, pitch, and duration). The values corresponding to step 1 (*DEsert*) and step 7 (*deSSERT*) in each dimension were based on the naturally produced tokens. The voiced portion of the syllables had its formant structure (F1, F2, F3, and corresponding bandwidths), duration, and pitch

manipulated, and intensity between the two stress patterns neutralized. Two of the dimensions (formant structure and pitch, formant structure and duration, pitch and duration) were orthogonally manipulated in 7 steps, holding the other two dimensions at Step 4 (for details, see [14]). This manipulation yielded 147 auditory stimuli (3 matrices of 7 × 7 stimuli).

2.3. Procedures

The complete experimental procedure lasted ten days. Participants completed the pre-test (i.e., the cue-weighting speech perception task) in the lab on the first day of participation. In each trial of the pre-test, participants heard an auditory stimulus over headphones and were asked to press the left arrow on the keyboard if they thought they heard *DEsert* and the right arrow if they thought they heard *deSSERT*. Each trial ended with the participant's response followed by a 1,000 ms inter-trial interval. The main session—including a total of 441 trials (147 stimuli × 3 repetitions)—was divided into three blocks, and test items were randomized across participants. The task lasted approximately 15 minutes.

The training was conducted remotely. On eight consecutive days following the pre-test, participants completed eight 20-minute training sessions, with no more than one training session per day. On each trial of the training, participants heard an auditory stimulus and were asked to decide whether the stimulus was a noun or a verb (they were told nouns would be stressed word-initially verbs word-finally), and they received explicit feedback on the accuracy of their responses. In each training session, participants heard 36 noun-verb pairs repeated 4 times (total: 288 stimuli). For the HVPT group, in each of Sessions 1-4, participants heard 2 talkers, with the stimuli being repeated twice per talker, and in each of Sessions 5-8, participants heard all 4 talkers. In the STT training, where participants heard only 1 talker, the stimuli were repeated four times. In each training session, 41.7% of the words had a H*L– intonation, 16.6% had an L*H– intonation, and 41.7% had a flat intonational contour, mimicking the distribution of word-level intonations reported in Im, Cole, and Baumann [18]. The distribution of intonational patterns was counterbalanced across talkers in the HVPT training and across the two stress patterns in both types of training.

On the day following the last training session, participants completed the post-test, which was identical to the pre-test.

2.4. Data analysis

Mixed-effects logistic regression models were conducted on Seoul Korean listeners' proportion of *DEsert* (coded as 1) and *deSSERT* (coded as 0). Separate models were built for each of the three stimulus matrices (i.e., vowel quality by pitch, vowel quality by duration, and pitch by duration). For each model, the fixed effects included two manipulated dimensions (each centered), their interactions, test (pre- vs. post-), and training type (HVPT vs. STT). Korean listeners' response on the pre-test in the HVPT condition served as baseline. Random intercepts were participant and item. The largest model was backward fit using log-likelihood ratios. Only the models with the best fit are presented.

3. RESULTS

Seoul Korean listeners' proportion of *DEsert* selection for each of the three stimulus matrices, test, and training type is shown in Fig. 1., together with English listeners' proportion of *DEsert* selection (for reference).

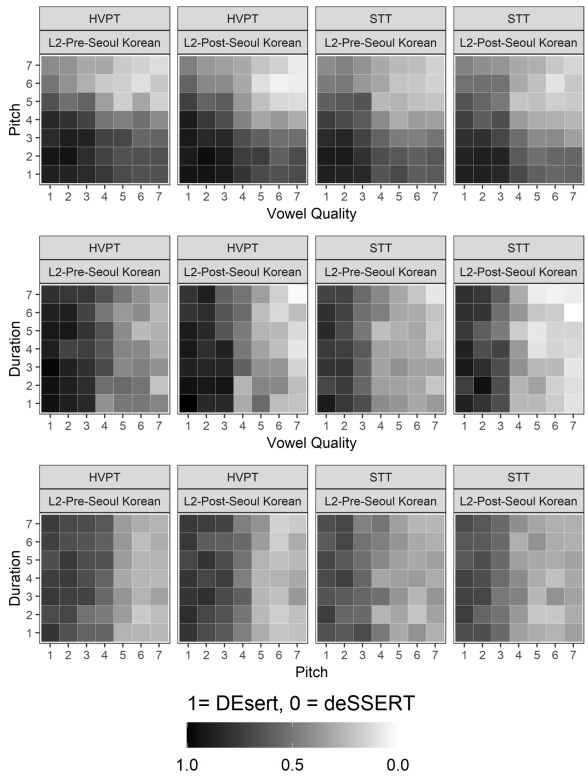

Figure 1: Seoul Korean listeners' proportion of *DEsert* vs. *deSSERT* selection as a function of test and training type when the stimuli varied by vowel quality and pitch (top), vowel quality and duration (middle), and pitch and duration (bottom). The two left panels show the results of HVPT, and the two right panels show the results of STT.

When the stimuli varied by vowel quality and pitch (top panels), the model with the best fit (Table 1) had the following structure: responses ~ (vowel.quality + pitch) * test + (vowel.quality + pitch) * training.type + (1|participant) + (1|item). The model revealed

significant effects of vowel quality, pitch, and test. Importantly, the model yielded a two-way interaction between vowel quality and test, with Korean listeners in the HVPT group showing a stronger effect of vowel quality (decrease in *DEsert* selection as step increased) in the post-test than in the pre-test, and an interaction between pitch and training type, with the effect of pitch (decrease in *DEsert* selection as step increased) being stronger in the pre-test results of the STT group than in those of the HVPT group. The lack of three-way interaction between cue, test, and training type suggests that HVPT is not superior to STT for altering listeners' cue weighting.

Table 1. Mixed-effects Logistic Regression with Best Fit for Vowel Quality (VQ) x Pitch Stimuli

	Est.	SE	*z*	*Pr*(>|*z*|)
(Intercept)	0.25	0.09	2.77	.006
VQ	−0.39	0.03	−13.02	< .001
Pitch	−0.54	0.03	−18.30	< .001
Test (Post-test)	−0.11	0.04	−3.10	.002
Training (STT)	−0.05	0.10	<|1|	>.1
VQ × Test (Post-test)	−0.06	0.02	−2.86	.004
Pitch × Test (Post-Test)	−0.03	0.02	−1.34	>.1
VQ × Training (STT)	0.01	0.02	<|1|	>.1
Pitch × Training (STT)	0.12	0.02	6.17	< .001

When the stimuli varied by vowel quality and duration (middle panels), the model with the best fit had the following structure: response ~ (vowel.quality + duration) * test + (vowel.quality + duration) * training.type + (1|participant) + (1|item). As can be seen in Table 2, the model revealed significant effects of vowel quality, duration, and test. Again, the model yielded a significant two-way interaction between vowel quality and test, with Korean listeners in the HVPT condition showing a greater effect of vowel quality in the post-test than in the pre-test. No other interaction was significant, suggesting again that the two training types did not differ in their ability to alter listeners' cue weightings.

When the stimuli varied by pitch and duration (bottom panels), the model with the best fit had the following structure: response ~ (pitch + duration) * test * training.type + (1|participant) + (1|item). As presented in Table 3, the model revealed significant effects of pitch, duration, and test. Crucially, the model also yielded a significant two-way interaction between pitch and test, indicating that Korean listeners in the HVPT condition relied more on pitch from pre-test to post-test, and a significant pitch-by-training-type interaction, suggesting that Korean listeners' use of pitch in the pre-test was stronger for the HVPT group than for the STT group. Again, the lack of three-way interaction between cue, test, and training type indicates that the two training types did not differ in their altering of listeners' cue weighting.

Table 2. Mixed-effects Logistic Regression with Best Fit for Vowel Quality (VQ) x Duration Stimuli

	Est.	SE	*z*	Pr(>|*z*|)
(Intercept)	0.47	0.14	3.29	.001
VQ	−0.42	0.02	−18.09	< .001
Duration	−0.05	0.02	−2.33	.020
Test (Post-test)	−0.33	0.04	−9.32	< .001
Training (STT)	−0.19	0.20	<|1|	>.1
VQ × Test (Post-test)	−0.08	0.02	−4.45	< .001
Duration × Test (Post-test)	−0.02	0.02	−1.26	>.1
VQ × Training (STT)	0.00	0.02	<|1|	>.1
Duration × Training (STT)	0.00	0.02	<|1|	>.1

Table 3. Mixed-effects Logistic Regression with Best Fit for Pitch x Duration Stimuli

	Est.	SE	*z*	Pr(>|*z*|)
(Intercept)	−0.02	0.09	−0.18	>.1
Pitch	−0.44	0.02	−19.95	< .001
Duration	0.04	0.02	2.00	.045
Test (Post-test)	−0.14	0.03	−4.09	< .001
Training (STT)	−0.03	0.12	<|1|	>.1
Pitch × Test (Post-test)	−0.06	0.02	−3.25	.001
Duration × Test (Post-test)	−0.02	0.02	−1.26	>.1
Pitch × Training (STT)	0.13	0.02	7.43	< .001
Duration × Training (STT)	−0.02	0.02	−1.31	>.1

4. DISCUSSION AND CONCLUSION

This study investigated whether HVPT can help Seoul Korean listeners rely less on pitch and more on vowel quality when perceiving English lexical stress, and whether the benefits of HVPT are superior to those of STT. The results showed comparable effects of training types on Korean listeners' use of vowel quality cues to English lexical stress. The similar training type effects are attributed to the considerable intonational variability that was introduced in the stimuli for both training types, resulting in greater ability for listeners in the STT group to extract vowel quality cues to English lexical stress. The increased use of pitch with training in the absence of vowel quality cues indicates that the greater occurrence of H* than of L* in the training may lead listeners to rely more on this cue after the training (especially since this was the cue in the pre- and post-test stimuli), suggesting that future training should make pitch cues to lexical stress unpredictable. This is the first study to show beneficial effects of perceptual training on listeners' weighting of acoustic cues to lexical stress.

5. REFERENCES

[1] Logan, J. S., Lively, S. E., & Pisoni, D. B. (1991). Training Japanese listeners to identify English /r/ and /l/: A first report. *J. Acoust. Soc. Am., 89*, 874-886.

[2] 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. *J. Acoust. Soc. Am., 94*, 1242-1255.

[3] Iverson, P., Hazan, V., & Bannister, K. (2005). Phonetic training with acoustic cue manipulations: A comparison of methods for teaching English /r/-/l/ to Japanese adults. *J. Acoust. Soc. Am., 118*, 3267-3278.

[4] Huensch, A., & Tremblay, A. (2015). Effects of perceptual phonetic training on the perception and production of second language syllable structure. *J. Phon., 52*, 105-120.

[5] Wang, Y., Spence, M. M., Jongman, A., & Sereno, J. A. (1999). Training American listeners to perceive Mandarin tones. *J. Acoust. Soc. Am., 106,* 3649-3658.

[6] Gay, T. (1978). Physiological and acoustic correlates of perceived stress. *Lg. and Speech, 21,* 347-353.

[7] Beckman, M. E., & Edwards, J. (1994). Articulatory evidence for differentiating stress categories. In P. A. Keating (Ed.), *Phonological structure and phonetic form: Papers in laboratory phonology III* (pp. 7-33). Cambridge: Cambridge University Press.

[8] Lehiste, I. (1970). *Suprasegmentals.* Cambridge: MIT Press.

[9] Beckman, M. E., & Pierrehumbert, J. (1986). Intonational structure in English and Japanese. *Phonology Yearbook, 3,* 255-310.

[10] Ladd, D. R. (2012). *Intonational Phonology.* Cambridge: Cambridge University Press.

[11] Jun, S.-A. (1998). The Accentual Phrase in the Korean prosodic hierarchy. *Phonology, 15,* 189-226.

[12] Lin, C. Y., Wang, M. I. N., Idsardi, W. J., & Xu, Y. I. (2014). Stress processing in Mandarin and Korean second language learners of English. Bilingualism: *Lg. and Cog., 17,* 316-346.

[13] Connell, K., Hüls, S., Martínez-García, M. T., Qin, Z., Shin, S., Yan, H., & Tremblay, A. (2018). English learners' use of segmental and suprasegmental cues to stress in lexical access: An eye-tracking study. *Lg. Learning, 68,* 635-668.

[14] Tremblay, A., Broersma, M., Zeng, Y., Kim, H., Lee, J., & Shin, S. (2021). Dutch listeners' perception of English lexical stress: A cue-weighting approach. *J. Acoust. Soc. Am., 149,* 3703-3714.

[15] Lemhöfer, K., & Broersma, M. (2012). Introducing LexTALE: A quick and valid Lexical Test for Advanced Learners of English. *Behav. Res. Methods, 44,* 325-343.

[16] Cutler, A., & Carter, D. M. (1987). The predominance of strong initial syllables in the English vocabulary. *Computer Speech and Lg., 2,* 133-142.

[17] Davies, M. (2008-) The Corpus of Contemporary American English (COCA): 560 million words, 1990-present. Available online at https://corpus.byu.edu/coca/.

[18] Im, S., Cole, J., & Baumann, S. (2018). The probabilistic relationship between pitch accents and information status in public speech. *Proceedings of 9th International Conference on Speech Prosody* (pp. 508-511). Poznán, Poland.