KOREAN LISTENERS' WEIGHTING AND INTEGRATION OF CUES TO THE THREE-WAY KOREAN STOP CONTRAST

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

[1]University of Kansas, USA, [2]University of Texas at El Paso, USA, [3]Hanyang Institute for Phonetics and Cognitive Sciences of Language (HIPCS), Hanyang University, Korea
kimhj@ku.edu, actremblay@utep.edu, tcho@hanyang.ac.kr

ABSTRACT

The current study investigates whether listeners' cue weighting predicts their cue integration as the speech signal unfolds over time. It does so by testing the time course of acoustic cue integration in the processing of Korean stop contrasts by native Korean listeners. Listeners' weighting of voice onset time (VOT) and fundamental frequency (F0) to perceive Korean stop contrasts was measured with a cue-weighting speech perception task, and the time course of VOT and F0 integration was examined with a visual-world eye-tracking task. The results revealed that the timing of VOT integration is predicted by listeners' reliance on F0, with delayed integration of VOT in target-competitor pairs where F0 is a primary cue to process the stop contrast. These results suggest that listeners adapt their integration of acoustic cues in spoken word recognition based on the importance they assign to individual cues.

Keywords: Korean laryngeal stop contrast, speech perception, cue weighting, cue integration.

1. INTRODUCTION

Phonological contrasts in natural speech sounds are distinguished by multiple acoustic dimensions. The acoustic cues involved in a phonological contrast, however, do not have equal importance in speech processing. While some acoustic cues contribute more reliably to distinguishing phonological contrasts (primary cues), other cues are more variable (secondary cues) [1, 2].

An important issue in the processing of cues to phonological contrasts is that acoustic information from different cues often arrives asynchronously as the speech signal unfolds over time. One influential account for listeners' integration of asynchronous acoustic cues is the continuous integration approach [3, 4]. The approach assumes that each cue provides partial evidence for lexical access as soon as it arrives. This partial information results in the greater activation of lexical candidates that are consistent with it, and lexical activation is updated as the remaining cues become available over time. However, it is not clear how such cue integration that develops over the time course of speech processing is related to cue weighting and the primacy of available cues in perceiving speech segments.

The present study investigates the relationship between listeners' cue-weighting and cue integration: Are acoustic cues integrated continuously when the weight of the cue available later is greater than that of the cue available earlier? As a test case, we examined how individual native Seoul Korean listeners weight and integrate voice onset time (VOT) and onset fundamental frequency (F0) when hearing Korean words that begin with a laryngeal stop.

Korean has a three-way stop contrast, with VOT and onset F0 of the following vowel playing a role in distinguishing the contrasts. In word-initial position, the fortis stop has a short VOT and high F0, the lenis stop has an intermediate VOT and low F0, and the aspirated stop has a long VOT and high F0 [5]. In Seoul Korean, however, the VOTs of lenis and aspirated stops have merged over time, resulting in speakers being more likely to depend on the onset F0 difference of the following vowel when distinguishing lenis from aspirated stops [6, 7, 8, 9].

In the perception of the Korean stop contrasts, Seoul Korean listeners use VOT as a primary cue for perceiving the aspirated vs. fortis contrast but F0 as a primary cue for perceiving the aspirated vs. lenis contrast, and both VOT and F0 as primary cues for perceiving the lenis vs. fortis contrast [10]. These acoustic properties of the Korean stop contrasts suggest that it is important to study the time course of cue integration because they raise the question of whether the timing of cue integration is contingent on the primary cue being heard earlier (VOT) or later (F0) in the speech signal.

The present study used a cue-weighting speech perception task (Experiment I) to quantify listeners' relative weighting of cues to the stop contrast and a visual-world eye-tracking experiment (Experiment II) to test the time course of cue integration. Under the continuous integration account, two hypotheses can be made regarding the relationship between cue integration and cue weighting. One is that the time course of cue integration is associated with cue weighting (the *associated view*). Under this view, listeners would not wait for F0 to integrate VOT if the earlier cue (VOT) is a primary cue (aspirated vs.

fortis), but they would wait for F0 to integrate VOT if the later cue (F0) is a primary cue (aspirated vs. lenis; fortis vs. lenis). The alternative hypothesis is that cue weighting is not associated with the time course of cue integration (the *independent view*). This view predicts that listeners would not wait for F0 to integrate VOT regardless of whether the earlier cue (VOT) is a primary or secondary cue to the contrast.

2. METHODS

2.1. Participants

The participants were 62 native Seoul Korean listeners (38 female, mean age: 24.1, SD: 3.2). They were tested in Korea and received monetary compensation for their participation. All participants had normal or corrected-to-normal vision and no speech and hearing disorder history.

2.2. Experiment I

2.2.1. Materials

The auditory stimuli for Experiment 1 were based on a Korean triplet *ppul* [p*ul] 'horn', *bul* [pul] 'fire', and *phul* [p^{h}ul] 'grass' [10]. The triplet was recorded by a female Seoul Korean speaker. The VOT of the word-initial stop and the onset F0 were each manipulated along a seven-step continuum. The minimum and maximum values of the VOT (0-110 ms) and F0 (180-290 Hz) continua were selected based on the acoustic analysis of the recorded tokens. The manipulation resulted in 49 auditory stimuli (7 steps of VOT × 7 steps of F0). All stimuli were normalized to a mean amplitude of 70 dB.

2.2.2. Procedure

The complete stimulus matrix was heard three times, once for each contrast pair (i.e., fortis vs. lenis choice, lenis vs. aspirated choice, and fortis vs. aspirated choice). On each trial, participants were asked to select the word that best matched what they heard by pressing either the left or the right arrow on the keyboard. The word labels were given with corresponding buttons on the computer monitor, and the position of the word labels was counterbalanced across trials. A total of 441 trials (49 stimuli × 3 pairs of choices × 3 repetitions) were randomly presented in three blocks, with each stimulus being heard once per block.

2.2.3. Data analysis

Mixed-effects logistic regression analyses were conducted on the participants' responses in each trial. The first model focused on listeners' responses to the aspirated vs. lenis contrast (aspirated response = 1). The fixed effects included VOT, F0, and their interaction. The initial model included random intercepts for item, participant, and repetition, and random slopes of VOT and F0 for each participant. The largest model was backward fit, and the best model was selected using the log-likelihood ratio comparisons. The second and third models focused on the lenis vs. fortis contrast (lenis response = 1) and the aspirated vs. fortis contrast (aspirated response = 1), respectively. When two cues showed significant effects, the fixed-effect coefficients were compared to determine which cue had a stronger effect than the other cue [11].

2.3. Experiment II

2.3.1. Materials

Fifteen disyllabic Korean words (five disyllabic noun triplets) that begin with a bilabial stop were selected as the critical stimuli. The words within a triplet share the same syllable structure and phonemes in the first syllable except for the word-initial stop, such that there is a temporary lexical ambiguity contingent on the initial consonant (e.g., *ppaltae* 'straw' - *balgul* 'digging' - *paljji* 'bracelet'). The critical words differed further at the onset of the second syllable. The words were controlled for (log-transformed) token frequency and number of letters.

The words in the display were presented orthographically in pairs. Along with the critical stimuli, there were 15 filler stimuli (five disyllabic triplets) with a three-way affricate contrast in word-initial position. Like the critical stimuli, the words within a filler triplet shared the same syllable structure and phonemes in the first syllable except for the word-initial affricate.

Critical trials consisted of six target-competitor types (3 target types × 2 competitor types): the fortis target and lenis competitor; the fortis target and aspirated competitor; the lenis target and fortis competitor; the lenis target and aspirated competitor; the aspirated target and fortis competitor; and the aspirated target and lenis competitor. On each trial, a target-competitor pair was grouped with one filler pair in the visual display, making a quadruplet (e.g., *ppaltae* 'straw' - *paljji* 'bracelet', *jjokji* 'note' - *chokgam* 'feel').

The auditory stimuli were recorded by a female native Seoul Korean speaker. The speaker produced each word three times in the carrier sentence *jigeum __(l)eul nuleusejo* 'Now click on __'. One recording of the carrier phrase was selected for the experiment. The target words were elicited from their original

carrier sentence. Unlike Experiment 1, the VOT of the word-initial stop and onset F0 of the following vowel were manipulated to have specific VOT and onset F0 values (fortis stop: 0 ms VOT and 250 Hz F0; lenis stop: 55 ms VOT and 180 Hz F0; aspirated stop: 110 ms VOT and 290 Hz F0); these values fell within the natural distribution of values for the corresponding stops [10], and they ensured that VOT and F0 would arrive as asynchronously as possible.

Since manipulating VOT yields different lengths of the first syllable within a triplet and results in longer syllables, thus taking slightly more time to reach the point of disambiguation, the duration of the following vowel was manipulated such that the syllable would have the same duration across tokens within a triplet. All stimuli were normalized to a mean amplitude of 70 dB.

2.3.2. Procedure

Eye movements were recorded with a head-mounted SR EyeLink II eye-tracker, sampling at 250 Hz. The experiment began with calibrating the eye tracker, which was followed by four practice trials and the main session. In each trial, participants saw a display with four printed Korean words arrayed in an invisible 2 × 2 grid for 2,000 ms. The arrangements of the printed words in the grid were randomized on each trial. The preview was followed by a fixation cross in the center of the screen for 500 ms. As soon as the fixation cross disappeared, the four printed words appeared on display in the same location, and the auditory stimulus was given over headphones. Participants were asked to click on the target word corresponding to the auditory stimulus as quickly and accurately as possible.

There were 180 trials, including 90 critical trials (6 target-competitor types × 5 triplets × 3 repetitions) and 90 filler trials (6 target-competitor types × 5 triplets × 3 repetitions). These trials were presented in three blocks, each containing 60 trials, with randomized orders across participants.

2.3.3. Data analysis

For the eye movement data, the average proportion of fixations directed to each word was calculated in 4 millisecond time windows for each target-competitor type. Statistical analyses were conducted on the difference between the empirical log-transformed proportions of target and competitor fixations (i.e., target-over-competitor advantage). The time course of each target-competitor type was aligned with the VOT disambiguation point, which is defined as the point where listeners could begin using VOT to distinguish the target from the competitor. We used generalized additive mixed modeling (GAMM) to analyze the data due to the nonlinearity of the eye-fixation trajectories. Three binary smoothing models were built using the function *bam* of the R package *mgcv* [12]. Each model examined the time point from the VOT disambiguation onset where the target-over-competitor advantage becomes significantly above 0 (i.e., significant target-over-competitor advantage). The GAMM specification of each model included the difference between the non-linear patterns of a target-competitor pair, the non-linear difference over time in the general time pattern for each participant (random effects), and the by-listener linear random slope for the two target-competitor pairs.

3. RESULTS

3.1. Experiment I

Fig. 1 shows listeners' responses and fixed-effect coefficients extracted for each listener on each contrast type.

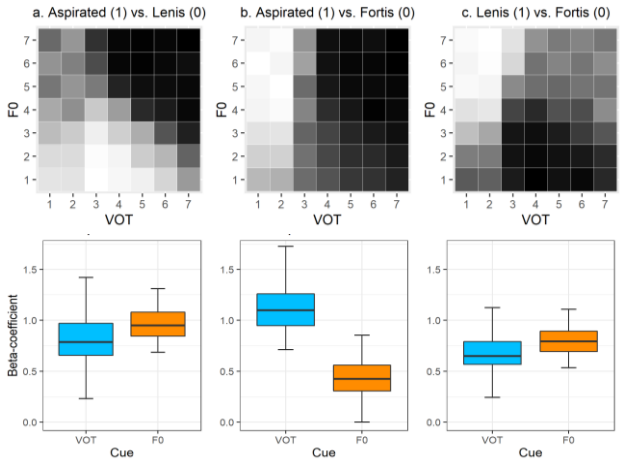

Figure 1: Listeners' responses (upper) and fixed-effect coefficients (lower) for the aspirated vs. lenis (left), the aspirated vs. fortis (middle), and the lenis vs. fortis (right) contrasts. The darker the cell, the more stop responses coded as 1.

For the aspirated vs. lenis contrast, the model found a significant simple effect of VOT ($\beta = 0.67$, $p < 0.001$) and F0 ($\beta = 0.94$, $p < 0.001$), with listeners' proportion of aspirated stop responses increasing as step number increased, and a significant two-way interaction between VOT and F0 ($\beta = 0.11$, $p < 0.001$), with the effect of VOT in Korean listeners' aspirated stop responses being greater at higher levels of F0 and with the effect of F0 being greater at higher levels of VOT. A comparison of cue importance revealed that the effect of F0 was stronger than that of VOT [$t(12535) = 2.92$, $p < 0.01$]. For the aspirated vs. fortis contrast, there was a significant simple

effect of VOT (β = 1.27, p < 0.001), with listeners' proportion of aspirated stop responses increasing as VOT step number increased, but no simple effect of F0. There was also a significant two-way interaction (β = 0.19, p < 0.001), with the effect of VOT in listeners' aspirated stop responses being greater at higher levels of F0 and the effect of F0 being greater at higher levels of VOT. For the lenis vs. fortis contrast, there was a significant simple effect of VOT (β = 0.56, p < 0.001) and F0 (β = −0.75, p < 0.001), with listeners' proportion of lenis stop responses increasing as VOT step number increased and F0 step number decreased. A comparison of cue importance showed that the effect of F0 was stronger than that of VOT [t(12497) = −17.87, p < 0.001].

These results suggest that Korean listeners rely more on onset F0 than on VOT to distinguish aspirated stops from lenis stops and lenis stops from fortis stops, and they rely more on VOT than on F0 to distinguish aspirated stops from fortis stops.

3.2. Experiment II

Fig. 2 shows listeners' target-over-competitor advantage in the target-competitor conditions.

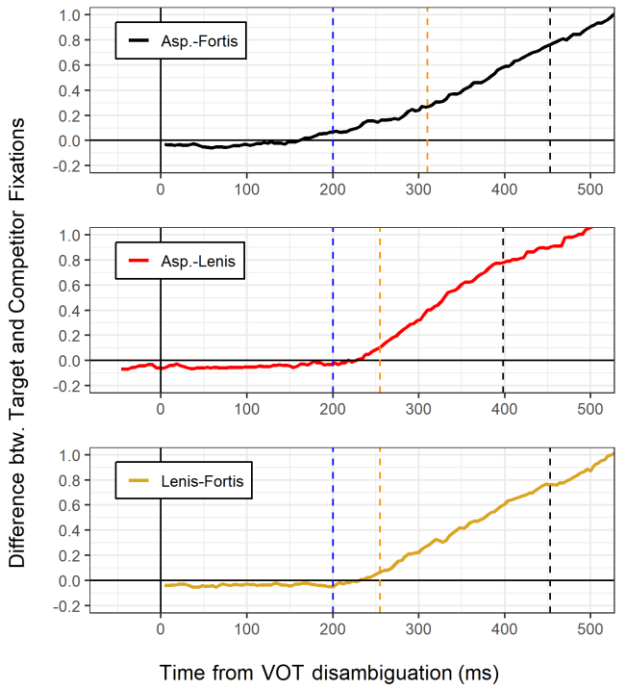

Figure 2: Listeners' eye movement in the condition with an aspirated target and a fortis competitor (top), an aspirated target and a lenis competitor (middle), and a lenis target and a fortis competitor (bottom). The vertical solid line represents the VOT disambiguation onset in the stimuli. The blue, yellow, and black dotted lines demarcate corrected VOT disambiguation onset, F0 onset, and second syllable onset (assuming a 200-ms oculomotor delay).

For the condition with an aspirated target and a fortis competitor (top panel), the GAMM analysis revealed that listeners start integrating VOT at 209 ms from the VOT disambiguation onset, suggesting that listeners showed a target-over-competitor advantage as soon as VOT became available in the speech signal. For the condition with an aspirated target and a lenis competitor (middle panel), the model showed that listeners started looking at the aspirated target at 248 ms from the VOT disambiguation onset, indicating a somewhat delayed lexical activation compared to the condition where the competitor began with a fortis stop. Similarly, for the condition with a lenis target and a fortis competitor (bottom panel), listeners started integrating VOT at 246 ms from the VOT disambiguation onset, suggesting that listeners' VOT integration was somewhat delayed.

These results indicate that listeners' real-time processing of acoustic cues for lexical access is modulated by their cue weighting, with delayed integration of VOT (early cue) in the target-competitor pairs where onset F0 (later cue) is a primary cue to process the stop contrast.

4. DISCUSSION

This study used a cue-weighting task and a visual-world eye-tracking paradigm to investigate how Seoul Korean listeners weight and integrate VOT and onset F0 to process Korean stop contrasts. The results support the *associated view* of cue integration, in that listeners' real-time processing of the cue available earlier in the speech signal was affected by the weight of the cue available later in the speech signal. These results suggest that the time course of cue integration is associated with the importance that listeners assign to different acoustic dimensions. From a theoretical perspective, the results elaborate more on the nature of the cue integration mechanism in spoken word recognition [3, 4] by providing the first empirical evidence for the effect of the relative perceptual weight of phonetic information on the time course of cue integration. Further research is needed to find converging evidence of the associated view of cue integration and to better understand the mechanism underlying the time course of cue integration, as it remains unclear whether cue integration is a consistent property of individual listeners or varies depending on the nature of acoustic cues involved in the phonological contrast.

Lastly, it is worth mentioning that a cue-weighting task showed that listeners do use VOT as a cue to the aspirated vs. lenis contrast, which is in line with the account discussed in [7] regarding the ongoing sound change in Seoul Korean lenis-aspirated distinction.

5. ACKNOWLEDGEMENTS

This work was supported by the KU Linguistics Graduate Student Research Scholarship.

6. REFERENCES

[1] Francis, A. L., Baldwin, K., Nusbaum, H. C. 2000. Effects of training on attention to acoustic cues. *Perception & Psychophysics* 62, 1668-1680.

[2] Holt, L. L., Lotto, A. J. 2006. Cue weighting in auditory categorization: implication for first and second language acquisition. *J. Acoust. Soc. Am*. 119, 3059-3071.

[3] McMurray, B., Clayards, M. A., Tanenhaus, M. K., Aslin, R. N. 2008. Tracking the time course of phonetic cue integration during spoken word recognition. *Psychonomic bulletin & review* 15, 1064-1071.

[4] Toscano, J. C., McMurray, B. 2010. Cue integration with categories: Weighting acoustic cues in speech using unsupervised learning and distributional statistics. *Cogn. Sci.* 34, 436-464.

[5] Lisker, L., Abramson, A. S. 1964. A cross-language study of voicing in initial stops: Acoustical measurements. *Word* 20, 384-422.

[6] Silva, D. J. 2006. Acoustic evidence for the emergence of tonal contrast in contemporary Korean. *Phonology* 23, 287-308.

[7] Choi, J., Kim, S., Cho, T. 2020. An apparent-time study of an ongoing sound change in Seoul Korean: A prosodic account. PloS ONE, 15 e0240682.

[8] Kang Y. 2014. Voice Onset Time merger and development of tonal contrast in Seoul Korean stops: a corpus study. *J. Phon.* 45, 77-90.

[9] Bang, H. Y., Sonderegger, M., Kang, Y., Clayards, M., Yoon, T. J. 2018. The emergence, progress, and impact of sound change in progress in Seoul Korean: implications for mechanisms of tonogenesis. *J. Phon.* 66, 120-144.

[10] 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. *J. Phon.* 41, 117-132.

[11] 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.

[12] Wood, S. 2006. Generalized additive models: an introduction with R. CRC Press.