

JOURNAL ARTICLE

The role of native phonology in spontaneous imitation: Evidence from Seoul Korean

Harim Kwon^{1,2}

¹ Department of Linguistics, University of Michigan, Ann Arbor, MI, US

² Department of English, George Mason University, Fairfax, VA, US

Corresponding author: Harim Kwon (hkwon20@gmu.edu)

This study investigates the role of phonology in spontaneous imitation in Seoul Korean speakers' imitation of aspirated stops by comparing the primary and non-primary cues. Seoul Korean aspirated stops are differentiated from stops of other phonation types by at least two distinct acoustic properties, stop VOT and f_0 of the post-stop vowel, with the latter being the primary cue. In the imitation experiment, Seoul Korean speakers heard and shadowed model speech that contained aspirated stops manipulated by either raising post-stop f_0 or lengthening VOT. Their realization of these properties in /t^h/, /t/, and /t*/ productions were compared before, during, and after exposure. Although both high f_0 and long VOT induced imitative changes in post-shadowing productions, the results revealed that exposure to an enhanced non-primary cue (long VOT) also influences the production of the primary cue for aspirated stops (post-stop f_0). However, an enhanced primary cue (high f_0) does not have similar effects on the non-primary cue. These results provide evidence that spontaneous imitation is not strictly tied to individual phonetic properties but it is rather phonological in that abstract categories are involved in the process of imitation.

Keywords: spontaneous imitation; phonological imitation; cue primacy; aspirated stop; Seoul Korean

1. Introduction

1.1. Spontaneous speech imitation

Spontaneous imitation (also known as convergence) refers to unintentional changes in speakers' productions in the direction of what they have recently heard. For instance, upon hearing model speech containing voiceless stops with longer voice onset time (VOT), English speakers spontaneously lengthen their own VOT of voiceless stops (e.g., Nielsen, 2011). This phenomenon provides evidence for a tight link between speech production and perception and, therefore, has been of great importance for phonetic theories that propose a strong relation between the two processes. In exemplar-based models of speech perception (e.g., Johnson, 1997, 2006; Pierrehumbert, 2001), the exemplars, or memory traces, activated in the course of perception potentially contribute to the subsequent productions, inducing imitative changes (e.g., Goldinger, 1998; Tilsen, 2009). Gestural theories of speech perception, particularly Direct Realism (Fowler, 1986, 1996), provide a rather different explanation for the mechanism underlying spontaneous imitation. It is claimed that vocal tract gestures are the common currency for speech perception and production. As listeners directly perceive articulatory gestures, perception can have an immediate impact on succeeding production (e.g., Fowler, Brown, Sabadini, & Weihing, 2003; Shockley, Sabadini, & Fowler, 2004). These two theoretical accounts are not

in direct opposition to each other, but they focus on different aspects of spontaneous imitation. While the exemplar account claims that the nature of the memory system gives rise to observed patterns of imitation with regard to the effects of lexical frequency, recency, and amount of exposure, the direct realist account claims that the intrinsic link between speech perception and production at the articulatory level explains the rapid and direct imitation.

Many researchers have investigated the role of various social factors in spontaneous imitation, and reported that non-grammatical or social factors, such as gender, conversational role, or attractiveness of the model speaker influence both the likelihood of imitation (i.e., whether the model speech is imitated in the subsequent productions or not) and the imitation fidelity (i.e., how similar the subsequent production becomes to the model speech) (e.g., Pardo, 2006; Babel, 2012; Babel, McGuire, Walters, & Nicholls, 2014). While these social aspects of imitation are an intriguing topic that has been intensively explored, the present study focuses on the role of phonology in imitation and asks the following question: When speakers converge to a model speaker, that is, when the imitation is triggered, how does the imitators' phonological knowledge influence the pattern of imitation (i.e., *how* the model is imitated), as well as the likelihood and the degree of imitation?

Previous research on the role of phonology in spontaneous imitation have suggested that changes in certain phonetic properties are not imitated when they are detrimental to phonological contrast or phonologically irrelevant. For example, Nielsen (2011) reports that English speakers imitate lengthened VOT for English voiceless /p/, but not shortened VOT. This suggests that the imitation is phonologically selective: Imitation is attenuated by the presence of a phonological boundary. In addition, Mitterer and Ernestus (2008) report that duration of pre-voicing of Dutch voiced stops is not imitated, attributing the lack of imitation to the phonological irrelevance of the property. According to Mitterer and Ernestus, in Dutch, presence versus absence of pre-voicing, but not its temporal extent, is phonologically relevant and, consequently, speakers do not imitate longer versus shorter pre-voicing.

Mitterer and Ernestus also examine whether Dutch speakers imitate different variants of Dutch /r/ (alveolar and uvular trills) and further claim that imitation occurs on an abstract phonological level. The two variants of Dutch /r/ are different articulatorily but equivalent phonologically, and the speakers hardly deviate from their habitual articulation to imitate the variant they have heard. Moreover, the response latency does not increase due to a gestural mismatch between the model speech and participants' response. Based on these findings, Mitterer and Ernestus claim that phonetic details, whether they are articulatory or acoustic/auditory, are not imitated when they are not phonologically relevant (see, however, Honorof, Weihing, & Fowler, 2011, p. 35).

However, the list of phonetic properties that have been reported to be susceptible to imitation is not limited to 'phonologically relevant' properties. For example, fundamental frequency (f_0) and duration of English vowels are arguably phonologically irrelevant but are found to be susceptible to imitation (e.g., Babel & Bulatov, 2012; Kim, 2012; Pardo, Gibbons, Suppes, & Krauss, 2012; Pardo, Jordan, Mallari, Scanlon, & Lewandowski, 2013). Furthermore, Honorof et al. (2011) point out that, if longer versus shorter pre-voicing of Dutch voiced stops is phonologically irrelevant, extended VOTs of English voiceless stops, one of the phonetic properties that have been repeatedly reported to be imitated in the literature, are arguably not relevant as well, because the lengthened VOTs (e.g., around 130 ms in Fowler et al., 2003, and 110 ms in Nielsen, 2011) are clearly within the aspirated allophonic category of English voiceless stops.

Mitterer and Müsseler (2013) provide an alternative interpretation of the non-imitation of duration of pre-voicing in Dutch: The difference between longer versus shorter pre-voiced stimuli used in Mitterer and Ernestus (2008) might have not been perceptually salient, precluding imitation. Some recent findings, including Mitterer and Müsseler's own, indeed provide (unsurprising) evidence that perceptually more salient variation induces more robust imitation. Mitterer and Müsseler examine the imitability of two different regional variations in German, and show that more marked dialectal variations lead to more robust imitation. Specifically, fricative-stop cluster variants [st~ʃt] induce more imitation than -ig variants [ɪk~iç] do, because the former is more marked, such that it clearly indexes a non-standard regional dialect, than the latter. Also, the imitation is greater when participants have heard both variants than when they have heard only one variant in an experiment, arguably because hearing both variants in the same experiment makes the variation more salient (Mitterer & Müsseler, 2013).

Although perceptual salience is a poorly defined term, it is widely accepted that a variant is more salient if it is less expected (hence more surprising) in a given context (see Schmid & Günther, 2016, for further discussion). Therefore, the typicality or naturalness of the variants in the model speech is expected to influence the degree of imitation. Indeed, model speech with less expected elements, such as words with low lexical frequency than those with high frequency (Goldinger, 1998; Goldinger & Azuma, 2004) and gender-atypical voices than typical ones (Babel et al., 2014), is reported to be more robustly imitated. Furthermore, when the model speaker is more deviant from the imitators' own phonetic repertoires, the large phonetic distance seems to result in more robust imitation (Babel, 2012; Walker & Campbell-Kipler, 2015).

Phonologically less-natural variants are also presumably more salient, often leading to more robust or longer-lasting imitation. For instance, Honorof et al. (2011) find that more velarized variants of /l/ in syllable onset induce greater imitation presumably because darker onset /l/ in American English is less typical with regard to allophonic rules. Furthermore, Zellou, Scarborough, and Nielsen (2016) investigate the imitation of coarticulatory vowel nasalization in English words from either high- or low-density phonological neighborhoods, and find that a less natural coarticulatory pattern—a decrease (versus an increase) in coarticulatory vowel nasality in English words from dense neighborhoods—induces longer-lasting imitation effects than a more natural pattern. Although both the increase and decrease in nasality were imitated during shadowing (i.e., immediate repetition of the model speech), only the imitation of decreased nasality persisted into a post-shadowing word-reading task.

Previous studies have also investigated the potential impacts of phonology on the process of spontaneous imitation through examining how imitation effects are generalized to unheard segments or words. In Nielsen (2011), after hearing English target words beginning with /p/ with extended VOT, participants produce extended VOT on unheard /p/-initial words and /k/-initial words as well as on the exposed target words. These productions indicate both phoneme-level generalization (to new stimuli including the same segment of exposure) and feature-level generalization (to a new segment that shares a feature), and point toward the influences of these abstract phonological units on the effects of imitation.

1.2. Seoul Korean voiceless stops

Korean has a three-way laryngeal contrast for voiceless stops: tense /p*, t*, k*/, lax /p, t, k/, and aspirated /p^h, t^h, k^h/. All three categories are phonetically voiceless in word- or phrase-initial positions although the lax stops are often voiced word- or phrase-medially.

In initial positions, contemporary Seoul Korean maintains this three-way contrast with at least two distinct acoustic cues, stop VOT and f_0 of the following vowel: Tense stops have short-lag VOT and high post-stop f_0 , lax stops have long-lag VOT and low f_0 , and aspirated stops have long-lag VOT and high f_0 . Studies on Seoul Korean stops suggest that post-stop high f_0 has become the primary cue for phonological *aspiration*,¹ as a consequence of a tonogenesis-like sound change that is in-progress (or recently completed) (e.g., Silva, 2006; Kang, 2014; among others). Earlier analyses of Korean stops describe tense stops as having short VOT and high f_0 , lax stops as having longer VOT and lower f_0 , and aspirated stops as having the longest VOT and higher f_0 (e.g., Kagaya, 1974). In contemporary Seoul Korean, however, the contrast between word-initial aspirated and lax stops is best differentiated by post-stop f_0 in production, replacing now-neutralized VOT difference between the two categories (e.g., Kang & Guion, 2008; Kang, 2014; Kong, Beckman, & Edwards, 2011; Lee & Jongman, 2012). Post-stop f_0 serves as a more reliable cue than VOT in perception as well (Kim, Beddor, & Horrocks, 2002; Kong et al., 2011; Lee, Politzer-Ahles, & Jongman, 2013).

As is usually the case for sound changes, female speakers have been reported to exhibit a more advanced stage than male speakers in this tonogenesis-like sound change of Seoul Korean. That is, the loss of the VOT distinction and development of a (pseudo-)tonal distinction between aspirated and lax stops hold true more for female speakers than male speakers (Kang, 2014; Oh, 2011). In addition, Kong et al. (2011) report that listeners were more sensitive to f_0 for the aspirated-lax distinction when they heard stops produced by female voices than those by male voices.

Due to physiological factors, f_0 at vowel onset is intrinsically correlated with the voicing of the preceding consonants (e.g., Hombert, Ohala, & Ewan, 1979). However, the high f_0 after aspirated stops of Seoul Korean is due to a language-specific phonological association of the acoustic property rather than a physiological consequence of long-lag VOT. When the increase in post-stop f_0 is a physiological epiphenomenon of long stop VOT, the change in f_0 is limited to the onset of the vowel (Hombert et al., 1979, among others). This is not the case in Seoul Korean, in which the f_0 difference between the aspirated and lax categories extends beyond the onset of the vowel (e.g., Kim, 2000; Kang, 2014). Moreover, VOT of long-lag voiceless stops decreases when produced in a higher f_0 range due to the increased tension in vocal folds (McCrea & Morris, 2005; Narayan & Bowden, 2013). This suggests that, in the case of Seoul Korean aspirated stops whose primary cue is high post-stop f_0 , the changes in VOT of the onset consonant and post-onset f_0 may not be positively correlated anymore.

In terms of phonological representations, two different feature systems have been provided. First, according to Halle and Stevens's (1971) binary feature-system, lax and aspirated stops share [+spread glottis] and tense and aspirated stops share [+stiff vocal cords]. Cho, Jun, and Ladefoged (2002), on the other hand, propose a privative feature system with underspecification: Aspirated and tense stops have [spread glottis] and [constricted glottis] respectively, while lax stops are unspecified. Relevant to the current study, the two feature systems² are crucially different in the sets of natural classes they provide. Under the binary system (Halle & Stevens, 1971), aspirated stops form a natural class with either lax stops or tense stops. Under the privative system (Cho et al., 2002), stops of different laryngeal categories do not form such natural classes. If imitation is generalized at the feature level, different natural classes predict different patterns of generalization. The specific predictions provided by different feature systems are presented in Section 1.3.

¹ Because the commonly used phonological label for Korean /p^h, t^h, k^h/ is aspirated stops, I will reserve the term *aspiration* or *aspirated* to refer to the abstract phonological category and use long VOT to refer to the acoustic property.

² Note both Halle and Stevens's (1971) and Cho et al.'s (2002) systems are based on rather earlier findings on Seoul Korean stops, when lax stops have intermediate VOT (longer than tense and shorter than aspirated stops).

1.3. Current study

The main goal of this study is to investigate the nature of the cognitive representations that are involved in spontaneous imitation, whether they are detailed phonetic properties (e.g., long VOT or high post-stop f_0) or phonological categories (e.g., stop aspiration). To this end, this study examines spontaneous imitation by separately manipulating two co-varying cues for one phonological contrast differing in their primacy. Seoul Korean speakers heard and shadowed model speech in which phonetic information for stop aspiration was manipulated. The model speech had word-initial /t^h/ either with raised post-stop f_0 or with extended VOT. The changes in participants' own productions were assessed by comparing the two acoustic properties of word-initial aspirated stops during shadowing of the model speech, as well as those before and after the shadowing block. In addition, unshadowed words beginning with stops of other laryngeal categories (i.e., /t/ and /t*/) as well as /t^h/ are compared before and after shadowing the model speech, to test the generalizability of imitative changes. The overarching question is whether cue primacy has an impact on spontaneous speech imitation. In order to investigate this broad question, this study asks several specific questions about the differences between the primary and secondary cues in spontaneous imitation.

The first question is which phonetic properties facilitate imitative changes when the model speech is manipulated such that either the primary or a secondary phonetic property for a phonological contrast is enhanced. In this study, the model speech includes aspirated stops with either the primary cue (post-stop f_0) or the secondary cue (stop VOT) enhanced. Note that both manipulations did not encroach on phonological boundaries (c.f., Nielsen, 2011), as they were in the direction of enhancing the phonological contrast between /t^h/ and the stops of other laryngeal categories (i.e., /t/ and /t*/). If enhanced phonetic properties cause imitative changes regardless of their cue primacy, both manipulations will induce some type of convergent changes. On the other hand, if only the primary cue for a phonological contrast instigates imitative changes, for speakers of Seoul Korean, only the enhanced post-stop f_0 will induce imitative changes relevant to stop aspiration. The existing literature is more consistent with the first hypothesis: Phonetic properties that arguably do not play a primarily contrastive role, such as English vowel duration (Kim, 2012; Pardo et al., 2012; Pardo et al., 2013) and vowel f_0 (Babel & Bulatov, 2012; Pardo et al., 2013) are spontaneously imitated. Based on these findings, I predict that enhanced phonetic properties, regardless of their cue primacy, will induce imitation.

One important caveat is that only those properties perceived by the listeners can facilitate imitation. If listeners do not detect (subconsciously) anything special or different about the model speech, they would not adjust (subconsciously) their subsequent productions based on what they have heard. To ensure that the manipulations are large enough to be perceived by the participants as being 'different,' a discrimination test is included in the current study. If participants do not perform better than chance in the discrimination test for the specific cue manipulation, no imitation is expected.

Assuming that the listener has detected the enhanced phonetic property, the next question is which phonetic property the listener-turned-speaker will adjust, if any, in her subsequent productions. For this question, two distinct hypotheses can be offered. The first possibility is that participants adjust the specific property that has been enhanced in the stimuli. That is, when the model speech has aspirated stops with long VOTs, participants will lengthen their VOTs for aspirated stops. Likewise, when the stimuli are aspirated stops with high post-stop f_0 , participants will raise their f_0 after aspirated stops. Crucially, under this hypothesis, the unmanipulated cue is not expected to change as a consequence of hearing the other cue enhancement. For instance, hearing high post-stop f_0 would not have an enhancing effect on stop VOT in the participants' productions, and neither would

long VOT on post-stop f_0 . Imitation, in this case, is strictly tied to a certain phonetic property, and thus I will refer to it as *phonetic* imitation.

The alternative hypothesis is that imitation is instead *phonological*, in which case, irrespective of the enhanced cue in the target stimuli, the listeners will enhance the property (or properties) they would normally use to enhance the relevant phonological category. For example, upon hearing aspirated $/t^h/$ with high post-stop f_0 or long VOT, participants perceive ‘enhanced aspirated stop’ and accordingly shift their production in that direction. This shift might involve both properties or, if a single property, might involve a property different from the one manipulated in the heard stimuli. This mechanism assumes that imitation is mediated by language-specific associations between phonetic properties and phonological categories. If imitation is phonological, the two manipulations are predicted to induce identical imitative patterns for a given speaker. That is, which phonetic property is enhanced in the stimuli does not matter as long as the participant detects the manipulation as enhancing, in this case, aspirated stops. The common imitative patterns are expected to involve enhancement of the phonetic property/properties the speakers would normally employ to enhance $/t^h/$. Young speakers of Seoul Korean increase post-stop f_0 to enhance aspirated stops (Kang & Guion, 2008), and thus I predict the participants in this study will adjust mostly post-stop f_0 in response to both manipulations. On the other hand, as younger female speakers exhibit the most advanced stage in the ongoing tonogenesis-like sound change in Seoul Korean (e.g., Kang, 2014), it is possible that at least some male participants may still rely on stop VOT at least to some extent. For these participants, if they exist, the post-stop f_0 is not the exclusive cue for stop aspiration, hence they are expected to adjust both properties.

Finally, this study asks whether the imitative changes are generalized to unheard words of different kinds. Three types of words that are not included in the model speech are compared before and after exposure to the model speech: $/t^h/$ -initial, $/t/$ -initial, and $/t^*/$ -initial words. The purpose is to test whether imitative behavior (whether it is to raise post-stop f_0 , lengthen VOT, or both) is generalized to unheard words sharing the target phoneme $/t^h/$, and unheard words with phonemes of different laryngeal categories $/t/$ and $/t^*/$. I predict that unheard $/t^h/$ -initial words will also show the imitative effect (phoneme-level generalization), replicating Nielsen’s (2011) findings.

As for the stops with different laryngeal categories, specific predictions depend on the feature system that is adopted. For instance, in Halle and Stevens’s (1971) system, aspirated and lax stops form a natural class sharing [+spread glottis], and tense and aspirated stops share [+stiff vocal cords]. Because the acoustic correlate of [+spread glottis] is long-lag VOT and that of [+stiff vocal cords] is high post-stop f_0 , this system predicts that the VOT of aspirated and lax stops shift together whereas the post-stop f_0 of aspirated and tense stops should shift in tandem. In contrast, under Cho et al.’s (2002) privative system, different laryngeal categories do not form such natural classes, and therefore lax and tense stops are not predicted to change as a consequence of hearing model speech with manipulated aspirated stops.

Note that the way the feature-level generalization is tested in this study is quite different from Nielsen (2011) as the relation between aspirated stops and lax stops in Seoul Korean may not be equivalent to that between English $/p/$ and $/k/$. The three stop categories under investigation here are distinguished by different laryngeal settings whereas English voiceless stops of different place of articulation share the same laryngeal settings. The current study explores whether imitative adjustments are extended to phonologically related categories when they are differentiated by the phonetic properties manipulated in the model speech.

2. Methods

2.1. Participants

Nineteen native speakers of Seoul Korean (12 female and 7 male, mean age = 25.2 years) participated. All participants were living in Ann Arbor, Michigan at the time of participation, and self-identified as native speakers of Seoul Korean. Three participants were born in the United States but returned to Korea before the age of five, and lived in Seoul for 15–23 years. The rest of the participants reported that they were born and raised in Seoul/Gyeonggi in Korea. All participants were proficient speakers of both Korean and English, but dominant in Korean. No participants reported any history of speech or hearing impairments. Each participant was paid \$25 for completing the two experimental sessions.

2.2. Stimuli

Korean words with initial /t^h/, /t/, or /t*/ were selected as test words from the NIKL corpus of modern Korean (morphologically parsed, corpus size = 15.3 million *eojeols*³) by the National Institute of Korean Language (2005). In addition, words with initial sonorants were selected as fillers. All test words were disyllabic, highly familiar (word familiarity scores being higher than 6.0 on a 7-point scale), and low in lexical frequency (below 50 in the NIKL corpus). The word familiarity score was obtained from ten native speakers of Korean who are different individuals from the participants of the main study. They were presented with all disyllabic words beginning with /t^h/, /t/, /t*/, /m/, /n/, /l/, /w/, and /j/ from the NIKL corpus, and asked to rate the familiarity of the words on a 7-point scale. Using the selected words, two wordlists (reading and shadowing lists) were constructed. The reading list contained 150 words: 50 /t^h-initial words, 25 /t/-initial words, 25 /t*/-initial words, and 50 sonorant-initial fillers. The shadowing list was a subset of the reading list (50 words), comprising half of the /t^h-initial words and half of the fillers from the reading list. For a complete list of stimuli, see Appendix A in the “Additional files” section.

A male native speaker of Seoul Korean (age = 25) served as the model speaker. Since younger speakers of Seoul Korean depend mainly on f₀ enhancement to enhance aspirated stops (Kang & Guion, 2008), and female speakers are in a more advanced stage than male speakers in the quasi-tonogenetic sound change in Seoul Korean (e.g., Kang, 2014; Kong et al., 2011), aspirated stops with lengthened VOT produced by a young female speaker can be perceived as ‘weird.’ Out of this concern, a male speaker was selected as the model speaker. The model speaker recorded the words from the shadowing list by producing the words in isolation three times in different randomized orders. He was instructed to speak naturally, at a normal speaking rate. His speech was digitally recorded onto a Macbook Pro laptop, using an AKG C 4000 B microphone and an external Edirol UA-25 preamplifier, with a sampling rate of 44.1 kHz via the Praat program (Boersma & Weenink, 2014). From the three repetitions, the best token of each item (free of unintended noises or mispronunciations) was selected for inclusion. All selected tokens were equalized to have an average intensity of 65 dB using the Scale intensity function in Praat.

The model speech for the targeted /t^h-initial words was manipulated in two ways using Praat. The high f₀ stimuli were created using the PSOLA method (Moulines & Charpentier, 1990), by raising the first pitch period of the post-/t^h/ vowel by 20% (calculated in Hertz value); f₀ of the rest of the first vowel was also raised proportionately. The long VOT stimuli were created by extending the VOT of word-initial /t^h/ by 60 ms. For each

³ An *eojeol* is an orthographic unit for Korean morphological analysis, which is separated by spaces. An *eojeol* can have one or more morphemes, or even words.

word, the medial portions of the phonetic aspiration were selected, copied, and pasted back into the aspiration section of the waveform. The duration of the selections as well as the number of splices varied across tokens in order to reach the target VOT duration without inducing any audible discontinuities. The mean VOT and post-stop f_0 (measured at the midpoint of the post-stop vowel) for the initial /t^h/s before manipulation were 58.38 ms and 153.6 Hz (7.6 semitone, henceforth St). The values after manipulation were 119.82 ms and 176.16 Hz (9.8 St), respectively. The mean post-sonorant f_0 for the filler words was 113.04 Hz (2.1 St).

2.3. Procedure

Each participant was tested in two experimental sessions that were conducted at least two weeks apart from each other. Each experimental session involved target stimuli with one of the two manipulations, raised f_0 or extended VOT. The order of the two experimental sessions was counterbalanced across participants to prevent any potential confounding effect of the testing order. Each session lasted approximately 30 minutes, consisting of an imitation experiment followed by an oddity discrimination test. Participants were also tested in two additional sessions involving English stimuli on different days for a separate study. On the last day of participation, after all other procedures, participants completed a questionnaire on their language background. The entire experiment was conducted in a sound-attenuated booth in the Phonetics Laboratory at the University of Michigan. All stimulus presentation was implemented using SuperLab stimulus presentation software (version 4.0.8, Cedrus Corporation) on a MacBook Pro laptop, with auditory stimuli being presented over AKG K271 MK II headphones.

2.3.1. Imitation experiment

The imitation experiment, using a slightly modified version of the word-naming paradigm (e.g., Babel, 2012; Goldinger, 1998; Nielsen, 2011), consisted of warm-up, baseline production, shadowing, and test production blocks. In the warm-up block, the words from the reading list were visually presented on the laptop screen, and the participants were asked to read them silently without pronouncing them. Each word was presented in the middle of the screen in Korean alphabet *Hangeul* one at a time, every 2 seconds, in a randomized order. In the baseline production block, the words were presented in the same way, but in a different random order. This time, the participants were instructed to read the words they saw on the screen aloud as clearly and promptly as possible. In the shadowing block, the words in the shadowing list with either the f_0 or VOT manipulation were played with nothing presented visually on the screen. The shadowing list was repeated three times, each time in different random orders without any break between repetitions. The participants were instructed to say aloud what they heard as clearly and promptly as possible. They were not instructed to imitate the stimuli. The inter-stimulus interval was 1.5 seconds. Finally, after the shadowing block, the test production block was conducted in the same fashion as the baseline production. Participants were allowed to take a short break between blocks, but no one rested more than a minute. All instructions during the experiment were given in Korean. The participants' baseline, shadowing, and test productions were digitally recorded onto a separate MacBook Pro laptop, using an AKG C 4000 B microphone and an external Edirol UA-25 preamplifier, with a sampling rate of 44.1 kHz via the Praat program (Boersma & Weenink, 2014).

2.3.2. Discrimination test

After completing each imitation experiment, the participants performed an oddity discrimination task that tested perception of the cue manipulation (stop VOT or post-stop f_0) used in the imitation experiment of that visit. The purpose was to determine whether the difference between the manipulated stimuli and the original recording was reliably perceived.

For each cue, 100 triplets were created from the same manipulated tokens that were used for the shadowing block in the imitation experiment and their original unmanipulated counterparts. Each triplet consisted of two identical tokens and an odd one. Half of the odd tokens were the manipulated ones, and the other half were the original ones. The place of the odd one in each triplet was decided pseudo-randomly, with the odd one appearing a roughly equal number of times in each of the three possible positions in each triplet. The interval between tokens within each triplet was 500 ms. One hundred triplets were presented in a randomized order over two experimental blocks with a self-paced break between blocks. Participants were asked to choose the odd one from the triplet by pressing a corresponding button on the button box (model RB-740, Cedrus Corporation). Each new triplet was played one second after the participant hit the button for the previous item. No feedback was provided during the test.

2.4. Measurements

Prior to making measurements, tokens with disfluency (0.8% of the total productions) were discarded. Disfluency was defined as when participants did not utter a word that they read or heard, said a different word, repeated a part of a word (including self-correction), or had some extra-verbal interruption such as coughing or clearing the throat. For f_0 analyses, an additional 17.1% of the total productions were excluded because the first vowels of the tokens were creaky voiced or completely voiceless. Creaky vowels occurred most commonly after tense $/t^*/$. Complete devoicing occurred frequently for $/i/$ following $/t^h/$ (e.g., 특가 $/t^hikka/$, 특진 $/t^hiktɕin/$).

The durations of the VOT of word-initial stop, the first vowel, and the word were measured for the remaining tokens by the author and a phonetically-trained research assistant. VOT was measured from the beginning of the release burst to the beginning of glottal pulsing in the waveform and/or the appearance of a voicing bar in the spectrogram. Prior to making the measurements, the file name of each recording was coded so that neither the author nor the assistant would know which production block (baseline, shadowing, or test) and manipulation condition (VOT extension or f_0 raising) the recording belongs to. A subset (10.8%) of the measurements was randomly chosen and analyzed to determine inter-rater consistency. The consistency score was computed using the Intraclass Correlation Coefficient (ICC) in the *irr* package (Gamer, Lemon, Fellows, & Singh, 2012) for R (R Development Core Team, 2014). All duration measurements were highly consistent between two raters (VOT ICC = 0.997; first vowel duration ICC = 0.989; and word duration ICC = 0.992; all p 's < 0.001).

At the temporal midpoint of the first vowel of each word, post-stop f_0 was measured using the pitch tracking function in Praat. f_0 measurements for all tokens were checked for tracking errors, and those with f_0 doubling or halving errors were hand-corrected. For subsequent statistical analyses, f_0 values were converted from Hertz to semitones.

3. Results

The participants performed the imitation task before the discrimination task out of the concern that the latter task, in which manipulated and original tokens were heard side by side, may have unwanted influence on the results of the imitation experiment. Despite the order of testing, here I report the results of the discrimination test first because they will work as the premise for the imitation experiment.

3.1. Discrimination test

Participants' sensitivity to the manipulated stimuli in the oddity discrimination task was converted to d' scores using the `dprime.oddity` function in the *psyphy* package (Knoblauch, 2014) for R. An analysis of variance (ANOVA) was conducted on d' scores in order to determine whether the participants' performance differed in the two manipulation conditions. The within-subjects independent variable was Manipulation Condition (long VOT, high f_0) with Subjects included as a random factor. The ANOVA results showed that the effects of Manipulation Condition were not significant [$F(1, 18) = 0.503, p = 0.487$], which suggests the participants' performance was not significantly better in one condition than the other. Moreover, the participants' performance on average (65% accuracy rate and $d' = 2.3$ in the high f_0 condition; 67% and $d' = 2.4$ in the long VOT condition) was reliably above the 33% chance level (or null sensitivity $d' = 0$).

The size of the two manipulations is not directly comparable since they are on different dimensions, spectral and temporal. Nonetheless, these results confirmed that the two manipulations resulted in perceptible differences of similar extents. Certainly, good performance in these oddity discrimination tests in which the manipulated stimuli were juxtaposed with the original ones does not guarantee that the participants would have noticed high post-stop f_0 or long stop VOT in the imitation experiments. If the difference had not been detected in discrimination testing, however, it would have been unlikely to be noticed in the imitation experiments as well. Therefore, the results of these discrimination tests serve as a prerequisite for the imitation experiments.

3.2. Imitation experiment

This section presents the data from the imitation experiment in two separate aspects: one focusing on the baseline-test comparisons to investigate the imitative changes that the primary and secondary cues caused in the participants' test productions (Section 3.2.1.), and the other on the changes in the shadowed /t^h/ productions (Section 3.2.2.). These two analyses were done separately because shadowing productions cannot be compared directly to read-speech productions (baseline/test).⁴ All statistical modeling was conducted in R using *lme4* (Bates, Maechler, Bolker, & Walker, 2014) and *lmerTest* (Kuznetsova, Brockhoff, & Christensen, 2014) packages, and parameter specific p -values and degrees of freedom were obtained by using Satterthwaite approximation. If a predictor is significant in multiple interactions (or a main effect and interactions), only the highest-level interaction is reported in prose, along with the results of post-hoc testing. The post-hoc pairwise comparisons were conducted using Tukey's HSD tests implemented in *emmeans* package (Lenth, 2018). For the descriptive statistics for stop VOT and post-stop f_0 measurements in baseline, shadowing, and test productions, see Appendix B in the "Additional files" section.

⁴ I thank the associate editor for pointing this out to me.

3.2.1. Baseline-Test comparisons

The differences between test productions of aspirated /t^h/, lax /t/, and tense /t^{*}/ and their baseline counterparts in the two manipulation conditions were statistically analyzed by fitting a linear mixed effects model to each property, stop VOT (ms) and post-stop f0 (St), on the subset of the data including read-speech (baseline/test) productions. Both models included fixed effects of manipulation CONDITION (high f0, long VOT), PRODUCTION block (baseline, test), onset TYPE (aspirated, lax, tense), presence of EXPOSURE (unheard, shadowed), and speaker SEX (female, male). Interaction terms among fixed factors were included when their inclusion resulted in an improved model fit based on a likelihood ratio test ($p < 0.05$). As a result, the final VOT model included COND * PROD * TYPE, SEX * TYPE interaction terms and the final f0 model included COND * PROD * TYPE, COND * PROD * SEX, COND * TYPE * SEX, and PROD * TYPE * SEX interaction terms. For the f0 model, ‘sonorant’ was added as an extra level of the factor TYPE to make sure that the observed changes in post-stop f0 were not due to global pitch shift. For similar reasons, the remaining word duration (REST = total word duration – VOT) was entered in the VOT model as a fixed effect to verify that the changes in VOT were not due to global changes in speech rate. All fixed factors were coded using treatment (dummy) coding, with the reference level for the intercept being set to high f0 (COND), baseline (PROD), aspirated (TYPE), unheard (EXP), and female (SEX). Random-effect structure was specified maximally as long as it allowed the models to converge and was permitted by the structure of the data (Barr, Levy, Scheepers, & Tily, 2013). This included (1) by-speaker and by-word intercepts, (2) by-speaker random slopes for COND, PROD, TYPE, and EXP, (3) by-word random slopes for COND and PROD. Adding more random slopes for interactions led to non-convergence, and thus was discarded. The outputs of the two mixed effects models are given in Table C.1–2 in Appendix C in the “Additional files” section.

The VOT model revealed a significant three-way interaction COND * TYPE * PROD. The results of this interaction are plotted in **Figure 1**. Additional post-hoc Tukey testing showed that the VOT of /t^h/ in the test block was significantly longer than its baseline

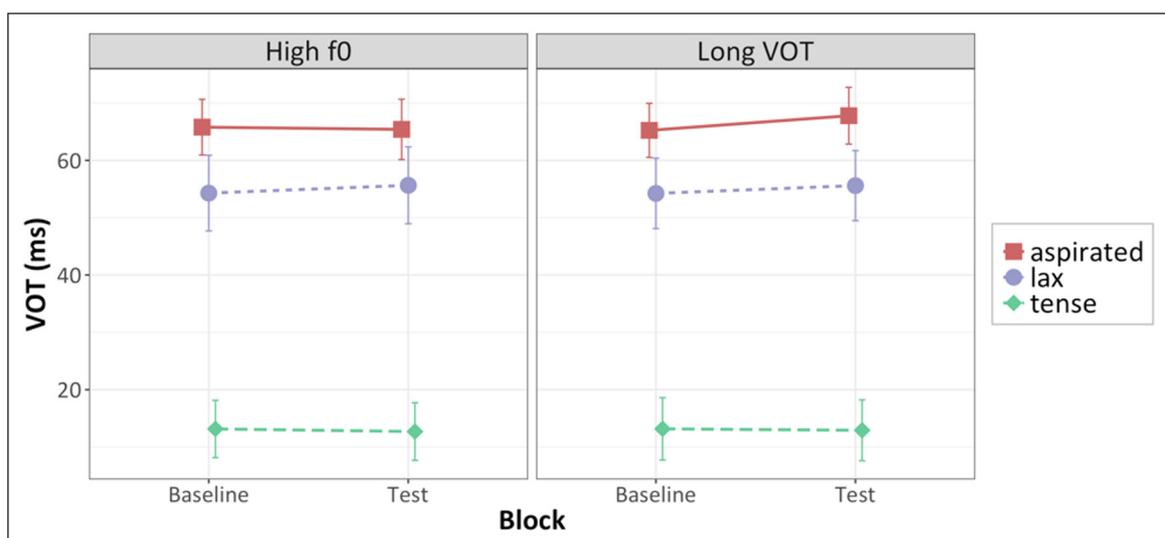


Figure 1: Estimated marginal means of stop VOT for different onset types in baseline and test productions of the two manipulation conditions (error bars indicate 95% confidence interval).

counterpart in the long VOT condition [$\beta = -2.559, t(45.35) = -3.248, p = 0.0022$]. On the other hand, the high f_0 manipulation had no effect on the VOT of /t^h/ [$\beta = 0.390, t(45.27) = 0.495, p = 0.6228$]. The VOTs of lax and tense stops were not different in baseline and test blocks in both manipulation conditions.

The VOT model also found a significant interaction of TYPE * SEX. The results of the post-hoc Tukey tests revealed that this interaction was due to the fact that the VOT difference between /t^h/ and /t/ was significant for the male participants [$\beta = 15.171, t(108.05) = 4.382, p < 0.0001$], but only marginal for the females [$\beta = 7.053, t(117.69) = 2.204, p = 0.0746$]. This is consistent with the previous findings that VOT of aspirated and lax stops overlaps more in female than in male productions (e.g., Kang, 2014; Oh, 2011). But this gender difference in production patterns did not seem to have led to different imitation behaviors (note that the interaction terms including SEX * COND were not included in the reported model because including them did not improve the model fit).

Another significant fixed effect in the VOT model was the effect of REST. This effect seems to be highly significant statistically [$\beta = -0.021, t(3068.0) = -7.635, p < 0.0001$], but the extremely small β value suggests that this effect is likely to be negligible in reality (VOT decreased by 0.021 ms when the rest of the word duration increased by 1 ms). Furthermore, the direction of the effect indicates that the changes in VOT and those in the rest of word duration are negatively correlated, if at all. This means that lengthening (or shortening) in VOT was not due to overall slowing down (or speeding up) of speech rate, and that the entire word duration was most likely to be constant despite changes in VOT. This result corroborates the previous findings that an increase in VOT is usually accompanied by a small decrease in the duration of the voiced portion of the post-stop vowel (e.g., Allen & Miller, 1999).

The f_0 model found significant three-way interactions for COND * PROD * TYPE, PROD * TYPE * SEX, COND * TYPE * SEX, and COND * PROD * SEX. These interactions are plotted in **Figure 2**. Because each of the four predictors was involved in three different

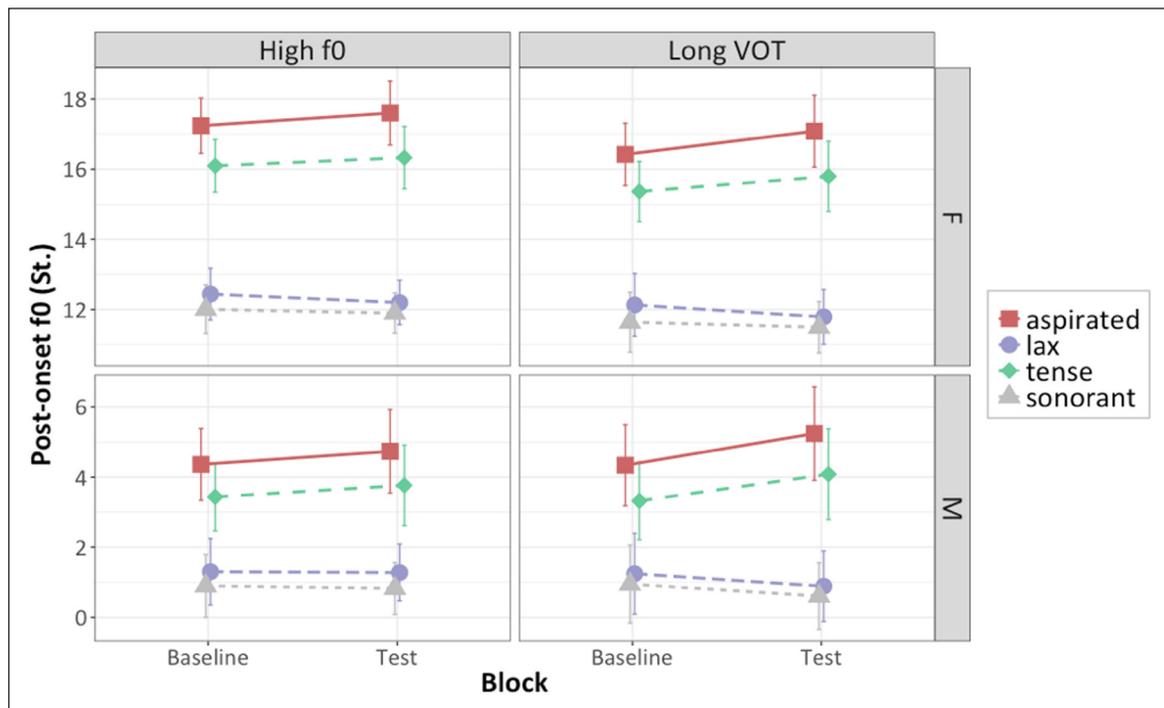


Figure 2: Estimated marginal means of post-onset f_0 for different onset types in baseline and test productions of the two manipulation conditions (error bars indicate 95% confidence interval).

interaction terms, these interactions were further explored by performing additional analyses on subsets of data split by onset TYPE. For the full outputs of these models, see Table C.3–6 in Appendix C in the “Additional files” section.

In the model that included only the aspirated stops, COND * PROD * SEX was significant. The post-hoc Tukey tests revealed the baseline-test difference was highly significant for all pairwise comparisons. Both female and male speakers produced significantly higher post-aspirated-stop f0 in the test block than in the baseline block in the high f0 condition [$\beta_{\text{FEMALE}} = -0.399$, $t(24.89) = -2.970$, $p = 0.0065$; $\beta_{\text{MALE}} = -0.310$, $t(24.92) = -2.783$, $p = 0.0112$], as well as in the long VOT condition [$\beta_{\text{FEMALE}} = -0.621$, $t(24.92) = -4.615$, $p = 0.0001$; $\beta_{\text{MALE}} = -0.976$, $t(24.95) = -5.468$, $p < 0.0001$].

This outcome indicates that, as can be seen in **Figure 2**, aspirated stops had higher post-stop f0 in test productions in both manipulation conditions, with greater between-block differences (baseline-test) in the long VOT condition than in the high f0 condition. This holds for both female and male speakers although the between-condition difference (higher f0-longer VOT) was greater for male speakers.

The tense stop model revealed a significant COND * PROD interaction. The post-hoc Tukey comparisons indicated that the baseline-test difference of post-tense-stop f0 was significant both in the high f0 condition [$\beta = -0.287$, $t(28.14) = -2.200$, $p = 0.0362$] and in the long VOT condition [$\beta = -0.592$, $t(28.05) = -4.546$, $p = 0.0001$]. The participants raised their post-tense-stop f0 after hearing and shadowing aspirated stops either with longer VOT or with enhanced f0. Unlike the aspirated stop model, this effect did not interact with SEX, suggesting that the effect held for both genders of participants.

The results of the lax stop model and the sonorant model were similar to each other. In both models, no interaction term was significant, and post-onset f0 was higher in the high f0 condition than in the long VOT condition [$\beta_{\text{LAX}} = -0.432$, $t(21.3) = -2.234$, $p = 0.0368$; $\beta_{\text{SONORANT}} = -0.443$, $t(20.0) = -2.386$, $p = 0.0271$]. This indicates that, regardless of the production block and speakers' gender, f0 following lax stops and sonorants was lower in the long VOT condition than in the high f0 condition. As this holds for both production blocks, it does not seem to be due to exposure to manipulated model speech. The baseline-test difference was not significant in the sonorant model [$\beta = 0.085$, $t(23.0) = 0.949$, $p = 0.3525$] and only marginally significant in the lax stop model [$\beta = 0.224$, $t(31.2) = 1.824$, $p = 0.0777$]. The direction of the marginal trend in the lax stop model suggests that participants might have lowered the post-lax-stop f0 after exposure to aspirated stops with f0 or VOT enhancements.

To summarize the results of these sub-analyses on post-onset f0, after exposure to aspirated stops with raised f0 or those with extended VOT, participants produced higher post-/t^h/ and -/t*/ f0 than their baseline. These observed increases in post-/t^h/ and -/t*/ f0 do not seem to be due to overall pitch raising, as the sonorant-initial fillers did not show increase in post-onset f0.

Finally, the fixed effect of EXP (shadowed vs. unheard) was significant neither in the VOT model nor in the f0 model, indicating that the shadowed words were not different from unheard words in this imitation task. Furthermore, the interaction terms including EXP * PROD were not included in both models because their inclusion did not improve the model fits. This suggests that regardless of whether the specific word was shadowed or not, the imitative change (or non-change) was not different. This result supports the claim that the target of speech imitation is not individual words (e.g., Nielsen, 2011).

3.2.2. Changes in /t^h/ during shadowing

Did the two manipulations differ in their impact on the shadowed /t^h/ productions? To answer this question, stop VOT (ms) and post-stop f0 (St) of the shadowed /t^h/s were statistically analyzed by fitting a linear mixed effects model to each property. Both models

included fixed effects of manipulation CONDition (high f0, long VOT) and speaker SEX (female, male), as well as two-way SEX * COND interactions. These models do not compare shadowing productions with the baseline counterparts because direct comparison between audio-prompted shadowing and text-prompted read-speech (baseline) can be misleading. In the VOT model, the remaining word duration (REST = total word duration – VOT) was included to verify whether the changes in VOT were due to global changes in speech rate. Similarly, to ensure that the observed changes in post-stop f0 were not due to global pitch shift that also influences post-sonorant f0, sonorant-initial fillers were included in the f0 model by including fixed effects of word TYPE (aspirated /t^h/-words, sonorant-initial fillers) as well as its interactions with other fixed terms (TYPE * SEX * COND). All fixed factors were coded using treatment (dummy) coding, with the reference level for the intercept being set to the high f0 (COND), female (SEX), and aspirated (TYPE). The random-effect structure included (1) by-speaker and by-word intercepts (in both VOT and f0 models), (2) by-speaker random slopes for COND and TYPE (only in the f0 model), and (3) by-word random slopes for COND (in both models). This was the maximal random-effect structure (Barr et al., 2013) that allowed the models to converge. The results of the two mixed effects models are reported in Appendix C (Table C.7–8) in the “Additional files” section.

In the VOT model, the COND * SEX interaction was significant (see **Figure 3**). The results of the post-hoc Tukey tests indicated that VOTs of shadowed /t^h/ in the long VOT condition were significant longer than those in the high f0 condition for both male and female participants while the female speakers had a greater mean difference than the males [$\beta_{\text{FEMALE}} = -6.349$, $t(59.78) = -9.728$, $p < 0.0001$; $\beta_{\text{MALE}} = -3.838$, $t(171.44) = -5.194$, $p < 0.0001$]. As in the baseline-test comparisons, the effect of REST was significant statistically, with a small negative β value [$\beta = -0.021$, $t(2064) = -3.434$, $p = 0.0006$]. This presumably suggests that the entire word duration was most likely to be constant despite changes in VOT.

The f0 model found a significant three-way interaction COND * TYPE * SEX (see **Figure 4**). The results of post-hoc Tukey tests indicated that female speakers had significantly higher post-/t^h/ f0 in the high f0 condition than in the long VOT condition [$\beta = 0.412$, $t(21.25) = 8.391$, $p < 0.0001$]. Crucially, these speakers' post-sonorant f0 was not different between the two manipulation conditions [$\beta = 0.378$, $t(21.26) = 1.610$,

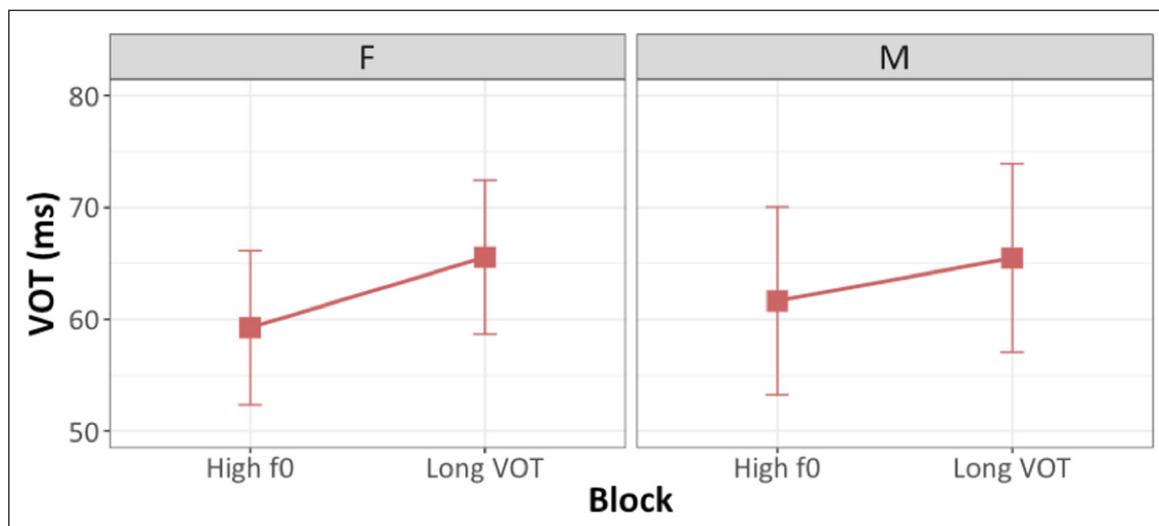


Figure 3: Estimated marginal means of stop VOT in shadowing productions of the two manipulation conditions (error bars indicate 95% confidence interval).

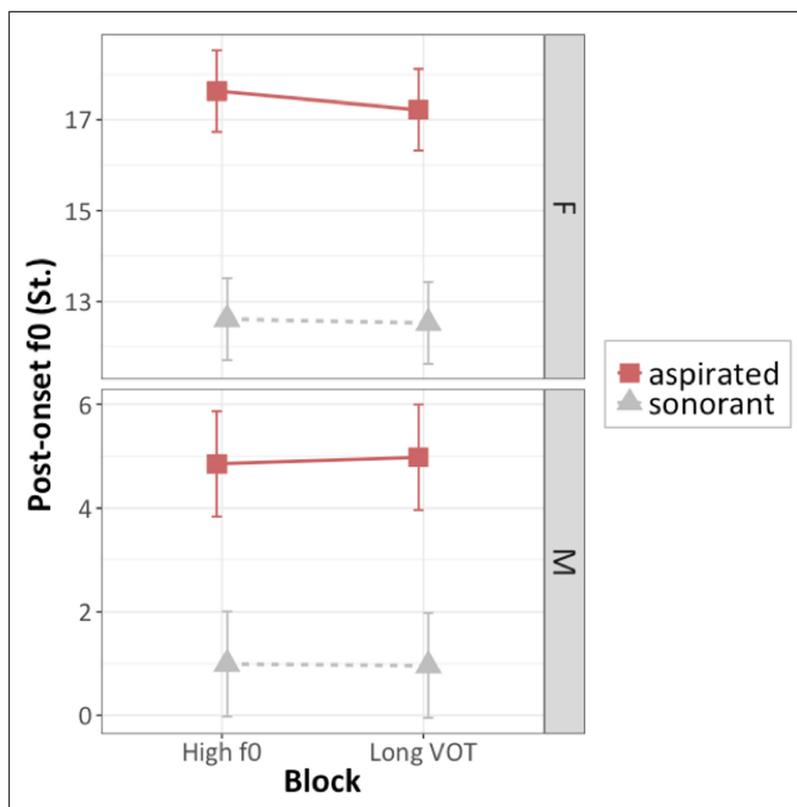


Figure 4: Estimated marginal means of post-onset f0 in shadowing productions of the two manipulation conditions (error bars indicate 95% confidence interval).

$p = 0.1222$], suggesting that the observed difference in post-/t^h/ f0 between the two manipulation conditions is not due to overall pitch raising or lowering. On the other hand, male speakers' post-onset f0 in shadowing was not different between the two conditions regardless of the onset types [$\beta_{\text{ASPIRATED}} = -0.124$, $t(21.24) = -0.290$, $p = 0.7746$; $\beta_{\text{SONORANT}} = 0.027$, $t(21.25) = 0.088$, $p = 0.9309$].

3.3. Summary of results

The current results reveal a number of patterns important to understanding the process of spontaneous imitation. First, in the discrimination test, participants reliably discriminated the manipulated stimuli from the unmanipulated originals regardless of specific manipulations. In addition, listeners' discriminability of the high f0 stimuli (primary cue enhancement) was not different from those of long VOT stimuli (non-primary cue enhancement).

These two different versions of stimuli led to distinct patterns of exposure-induced changes. The comparisons between baseline and post-shadowing test productions reveal that an enhanced non-primary cue for stop aspiration (long VOT) induced an increase in the primary cue whereas an enhanced primary cue (high post-stop f0) did not have comparable effects on the non-primary cue. That is, the Seoul Korean speakers *imitated* exaggerated stop aspiration cued by long VOT not just by lengthening the stop VOT but also by raising their f0 after those stops. After hearing aspirated stops with raised post-stop f0, however, the same participants only imitated the manipulated property (high post-stop f0) without lengthening VOT.

Exposure to the enhanced aspirated stops also caused a change in the production of the tense stop /t^{*}/. After hearing and shadowing /t^h/s enhanced either with VOT or post-stop f0, the participants' post-/t^{*}/ f0 increased compared to their baseline without notable changes in the VOT of /t^{*}/. The lax stop /t/ or sonorants did not change, except for a

marginal decrease in post-/t/ f₀. The changes in post-/t^h/ and -/t*/ f₀ were greater when the stimuli contained aspirated stops with lengthened VOT than when they contained those with raised post-/t^h/ f₀.

The two manipulations had a different impact on the participants' shadowing productions as well. Participants' own VOT of /t^h/ was longer when shadowing aspirated stops with longer VOT than when shadowing those with higher f₀. Note, however, that this does not imply that participants lengthened their VOT after hearing long VOT /t^h/ in comparison with their own baseline productions. The participants rather shortened their VOT when shadowing in the high f₀ condition than lengthened their VOT when shadowing long VOT stops (see Appendix B in the "Additional files" section). In addition to the difference in VOT, female speakers also produced higher post-/t^h/ f₀ when the model speech included aspirated stops with higher f₀ than when it included those with longer VOT. Male speakers did not produce different post-/t^h/ f₀ when shadowing /t^h/ with either manipulation.

The male and female speakers differed in post-/t^h/ f₀ changes in different production blocks in the two manipulation conditions. In test productions, male speakers showed greater post-/t^h/ f₀ difference between the two manipulation conditions (long VOT vs. high f₀) than females in test productions. However, only female speakers showed a between-condition difference in post-/t^h/ f₀ while shadowing the manipulated model speech.

Across all production blocks, stop VOT was negatively correlated with the rest of word duration. Shadowed and unheard /t^h-initial words were not different in their imitative changes.

4. Discussion

4.1. Cue primacy and spontaneous imitation

This study examined whether and how the primary and non-primary cues for stop aspiration exhibit different imitation patterns for speakers of Seoul Korean. Investigation of the effects of cue primacy on spontaneous imitation yielded a richer picture than has been presented in the literature as to the role of phonology in the process of imitation.

The first question posed in this study was which phonetic properties facilitate imitative changes, and it was predicted that both raised post-stop f₀ and extended VOT would induce imitative adjustments if participants perceive the manipulated stimuli as being different. The results were consistent with this prediction. The outcome of the discrimination tests confirmed that the participants reliably discriminated both types of the manipulated stimuli from the unmanipulated ones. And the two types of manipulated /t^h/ variants, one with the primary and the other with the secondary cues enhanced, triggered imitative adjustments in the participants' productions after exposure to the manipulations. (Note that increases in post-stop f₀ after having heard /t^h/ with long VOT are also referred to as imitative changes in this study.) This suggests that changes due to spontaneous imitation are facilitated by a phonetic property regardless of its primacy for a phonological contrast, as long as it is sufficiently perceptible.

Provided that the enhanced phonetic properties in the model speech were detected and induced imitation effects, the next question was as follows: Which phonetic property, if any, will the listeners-turned-speakers adjust in their subsequent productions? With regard to this question, two distinct hypotheses were offered. The phonetic imitation hypothesis predicted that the phonetic property manipulated in the model speech will match the property listener/speakers enhance in subsequent productions. The phonological imitation hypothesis predicted that, regardless of the cue manipulated in the model speech, listener/speakers will enhance the phonetic property/properties that they would normally use to enhance the relevant phonological category (in this study, aspirated stops).

The current outcome revealed intriguing asymmetries between primary and non-primary cues in imitation of Seoul Korean aspirated stops. First, in shadowing productions immediately after hearing the manipulated model speech, participants' /t^h/ productions were different in the two manipulation conditions, and the different productions seem to be roughly in line with the cue manipulated in the model speech. For instance, participants' shadowing /t^h/ VOT was longer in the long VOT condition than in the high f₀ condition, and female participants' shadowing f₀ was higher in the high f₀ condition than in the long VOT condition. These outcomes seemingly suggest that the participants, during shadowing, imitated the cue manipulated in the model speech phonetically. However, these need to be interpreted with caution, since the baseline and the shadowing productions were not directly compared in this study due to the different nature of the tasks (reading versus shadowing). As noted earlier, the VOT differences between the two manipulation conditions do not indicate that the participants lengthened (or shortened) their own VOT from the baseline counterparts. In addition, the baseline post-stop f₀ of female speakers was higher in the high f₀ condition than in the long VOT condition (see Appendix B in the "Additional files" section) so it is not conclusive whether the difference in their shadowing f₀ is due to imitative changes or a mere retainment of baseline pitch. Nevertheless, the participants produced quite different /t^h/s when shadowing different model speech, and the differences arguably show some correspondence to the acoustic targets that they shadowed.

In test productions, on the other hand, the imitative changes do not seem to be tied to the acoustic property manipulated in the model speech. Post-/t^h/ f₀, which is the primary cue for the relevant target phonological category (i.e., aspirated stops), increased significantly from the baseline block to the test block in both manipulation conditions. In addition, the test productions showed a significant increase in VOT only in the long VOT condition. This outcome seems to be incompatible with the phonetic imitation hypothesis as it cannot explain why participants raised post-/t^h/ f₀ after having heard /t^h/ with extended VOT.

Can this increase in post-/t^h/ f₀ in the long VOT condition be due to the physiological relation between stop VOT and post-stop f₀, such that long positive stop VOT leads to high post-stop f₀ (e.g., Hombert et al., 1979)? The consonant-induced changes in post-onset f₀ tend to be limited to the beginning of the vowel and small in size. However, in the current study, the changes in post-/t^h/ f₀, measured at the midpoint of the post-stop vowel, were greater in the long VOT condition (without the matching acoustic signals in the model) than in the high f₀ condition (see **Figure 2**). In addition, as mentioned in Section 1.2, when voiceless stops are produced in a higher f₀ range, their VOT decreases (McCrea & Morris, 2005; Narayan & Bowden, 2013), suggesting that the increase in f₀ may not arguably be a physiological consequence of lengthening VOT. Seoul Korean speakers' raising f₀ after having heard long VOT /t^h/ is due to language-specific phonological association rather than a physiological relation between the two properties. The participants in this study appear to have adjusted the cue that they would primarily use to enhance stop aspiration (i.e., post-stop f₀) in addition to imitating the cue that the stimuli they heard employed to enhance aspirated stops. This indicates that the target of speech imitation is not just the detailed acoustic parameters but rather abstract units (such as the phoneme /t^h/ or the natural class of aspirated stops, in this study). Exposure to an enhanced phonetic property can influence subsequent production not only of that property but also of other phonetic properties if they are important for the targeted phonological category.

However, the results raise one complication for the phonological imitation hypothesis, which predicts that participants will enhance the property/properties they would normally use to enhance the relevant phonological category, resulting in identical imitation patterns

irrespective of the enhanced cue in the model speech. The results appear to be inconsistent with this prediction. Specifically, the increase in the /t^h/ VOT was significant only in the long VOT condition, but not in the high f0 condition. Furthermore, the increase in participants' post-/t^h/ f0 was greater in the long VOT condition than the high f0 condition (see **Figure 2**), even when post-stop f0 in the high f0 stimuli was much higher than that in the long VOT stimuli. This difference arguably suggests that a mismatch with expectations about the voice may have made a variant more salient, resulting in more robust imitation. Since younger speakers of Seoul Korean depend mainly on f0 enhancement to enhance aspirated stops (Kang & Guion, 2008) and the model speaker of the current study is a young Seoul Korean speaker (age = 25), the participants presumably expected (subconsciously) that the model speaker would rely on f0 to enhance his /t^h/. When participants heard /t^h/ with extended VOT, that is, when the actual signal did not match with their expectation, the mismatch made the long VOT variants more salient,⁵ and thus facilitated greater changes in their own post-shadowing test productions. This corroborates previous findings that less expected variants, whether phonologically or socially, often lead to more robust imitation (e.g., Honorof et al., 2011; Mitterer & Müsseler, 2013; Babel et al., 2014; Zellou et al., 2016).

As for the extent to which imitative behavior (whether it is to raise post-stop f0, lengthen VOT, or both) is generalized, the current outcome first replicates the phoneme-level generalization in Nielsen (2011). The imitative changes in shadowed /t^h/-initial words (raising post-stop f0 and/or lengthening VOT) are generalized to unheard /t^h/-initial words in test productions, revealing no statistically significant difference between the shadowed and unheard words. In other words, hearing and shadowing specific words three times did not result in greater imitation of those words of exposure, calling into question the word-specificity of imitation suggested by Goldinger (1998).

This study also attempted to test whether the imitative enhancements are extended to stops with different laryngeal categories, namely, lax and tense stops. The current results show that post tense-stop f0 increased in both manipulation conditions: Otherwise, no other stop enhancement effects (increase in VOT or f0) were found. The increased post-/t*/ f0 may suggest that tense and aspirated stops share the feature whose acoustic correlate is high f0 (Halle & Stevens, 1971). The lack of change in lax /t/ indicates an apparent discrepancy between the current results and Nielsen's (2011) finding of feature-level generalization. This may be due to the fact that the relation between Seoul Korean /t^h/ and /t/ is not equivalent to that between English /p/ and /k/, despite the VOT merger between Seoul Korean /t/ and /t^h/. Furthermore, post-onset f0 of lax /t/ marginally decreased in test productions of both manipulation conditions. This tentatively suggests that the participants readjusted their productions in the direction that maximizes the relevant contrast (i.e., stop aspiration) by lowering post-lax-stop f0, although the current findings do not provide further evidence to verify this possibility. Further research would be needed to understand the interplay among the imitative behavior, its extension to phonologically relevant categories, and the cue primacy.

⁵ One of the reviewers suggested a post-hoc survey in order to verify whether the long VOT variant (but not the high f0 variant) is actually 'atypical' for speakers of Seoul Korean. I conducted a survey with a separate group of Korean listeners (n = 30) in which ten listeners heard one of the three variants of the /t^h/-initial stimuli used in this study (original, high f0, long VOT) and judged the model speaker's gender, age, and regional origin. All three variants were unanimously perceived as being produced by a male speaker, between 25–34 years old. As for the regional origin of the speaker, both original and high f0 stimuli were unanimously perceived as being produced by a Seoul Korean speaker. The listeners, however, disagreed on the long VOT stimuli. Out of ten listeners who heard the long VOT variant, two were unsure about the speaker's origin, another judged the speaker to be a foreigner, and the other seven to be a Seoulite. These results further support the possibility of the long VOT stimuli (but not the high f0 ones) being perceived as atypical.

4.2. Theoretical and practical implications

The results of this study provide evidence that the cognitive representations involved in the process of imitation, which bridges speech perception and production, draw on complex phonological categories such as stop aspiration rather than isolated acoustic properties, such as long VOT or high f_0 . Spontaneous imitation is phonological, rather than phonetic, such that it is governed by language-specific associations between phonological categories and acoustic properties. That is, the same acoustic property could have different phonological significance for speakers of different languages (as is well-known in cross-language speech perception literature) and, therefore, have different impact on the subsequent productions of speakers of different languages. Specifically, the phonological imitation observed in this study, a robust increase in post-stop f_0 after hearing /t^h/ with longer VOT, is not expected for speakers of a language in which f_0 is not a primary cue for the stop phonation types.

Because the imitative changes are not tied to a single acoustic cue, production patterns different from the model speech can emerge as a consequence of *imitation*. In other words, speakers' productions can converge to the ambient speech by becoming more different from it in a targeted dimension. This is in line with recent findings on the imitation of coarticulatory nasality (Zellou et al., 2016). Speakers do not imitate the actual degree of nasality but the abstract degree of nasality computed in comparison with oral fillers. The current findings provide further evidence for the claim that the target of imitation draws on abstract categories.

This has a practical implication for future studies on speech imitation. When perceptual judgments of listeners who are asked to assess imitated speech do not match the acoustic measurements (Pardo, 2013; Pardo et al., 2013), it could be that speakers are imitating phonologically. Phonological categories are signaled by multiple, often co-varying, phonetic properties, and researchers cannot always be sure which phonetic cues listeners will attend to and listeners-turned-speakers will adjust in the process of speech imitation. Therefore, future studies on speech imitation should consider doing more than relying solely on one phonetic measurement to assess imitation, especially when multiple co-varying cues are known to interact to signal one phonological category as in this study (see also Pardo, 2013).

How would phonological imitation be explained in different phonetic theories? First, the asymmetry between the primary and the non-primary cues in spontaneous imitation found in this study is consistent with the version of exemplar models that allows abstract linguistic levels (e.g., Pierrehumbert, 2001, among others). If a speaker of Seoul Korean hears instances of /t^h/ with consistently high f_0 and variably long VOT, her category of /t^h/ will have many high f_0 instances and fewer long VOT instances. When she hears an enhanced /t^h/, all the exemplars associated with the enhanced /t^h/ will be activated and contribute to the subsequent production. In such a system, the probability that a speaker would use high f_0 /t^h/ or long VOT /t^h/ variants is determined by the proportion of the two variants stored in memory and the associations among those traces. Because linguistic experiences themselves are the phonological categories, governing both the perception and production of speech, the role of the language-specific association between phonetic properties and phonological categories in imitation can be handled easily.

On the other hand, phonological imitation does not obviously follow from a direct realist account of imitation because listeners are not expected to imitate one gesture using a different gesture if they perceive gestures directly (Fowler et al., 2003; Honorof et al., 2011). This prediction of direct gestural perception and imitation remains the same regardless of how closely the two gestures are related in terms of phonology. Contrary to

this prediction, the current outcome shows that listeners supplement (or substitute) long VOT with high f_0 in spontaneous imitation, suggesting that abstract linguistic categories mediate what listeners perceive and what listeners-turned-speakers produce. Seoul Korean speakers, after perceiving a specific timing between the oral constriction gesture and the glottal opening-and-closing gesture that gives rise to longer VOT, adjust their laryngeal configuration to stiff vocal folds and narrower glottal opening because their phonological grammar specifies the association between the two phonetic properties.

However, phonological imitation is not a challenge to the direct realist account of imitation if it is accompanied by longer response latency. The direct perception and imitation of gestures (without an intervening abstract category) are predicted during rapid shadowing. Hypothetically, phonological imitation involving gestural supplement or substitution might take longer than direct (phonetic) imitation of perceived gestures. The current study was not designed to answer this question. Although the instruction for the shadowing block emphasized ‘quick’ responses, participants were not especially pressed for time; they could take as much time as they needed to shadow the word they heard within a 1.5 second inter-stimulus interval, which is relatively long for a shadowing task (c.f., the 180–250 ms latency lag for shadowing reported by Fowler et al., 2003). In fact, Mitterer and Ernestus (2008) reported that the mismatch between the shadower’s and model’s gestures of phonologically equivalent Dutch trills did not cause an increase in latency at the range of 200–600 ms latency lag. This is considerably longer than Fowler et al. (2003), as pointed out by Honorof et al. (2011), which may have led to their differing results. These previous discussions taken together, it is possible that speech gestures are directly perceived and imitated at rapid shadowing (e.g., within 200 ms response latency), but in the succeeding production with latency longer (e.g., ≥ 200 ms), the constellation of co-occurring gestures specified by the phonological grammar changes together. Further investigation will be needed to verify whether the difference in response latency actually results in different patterns (phonetic versus phonological) of imitation.

Finally, the current findings suggest that participants’ production and perception grammars would be related but may not be identical. Although Seoul Korean speakers rely primarily on post-stop f_0 to enhance / t^h / in their own speech, perceptually they must also be sensitive to VOT as information for / t^h / because otherwise they could not interpret extended VOT as a cue for enhanced / t^h /. This production-perception difference may not be surprising. As speech is highly variable (perhaps even more so with a recent sound change as in Seoul Korean), a speaker-listener’s perception grammar is necessarily more comprehensive than one’s own production grammar for effective communication. Speakers need to perceive distinctions among speech patterns that they do not normally produce themselves but others around them produce.

Additional Files

The additional files for this article can be found as follows:

- **Appendix A.** The list of stimuli. DOI: <https://doi.org/10.5334/labphon.83.s1>
- **Appendix B.** Descriptive statistics for stop VOT and post-stop f_0 in imitation. DOI: <https://doi.org/10.5334/labphon.83.s2>
- **Appendix C.** Statistical results – model outputs. DOI: <https://doi.org/10.5334/labphon.83.s3>

Acknowledgements

Portions of this work appear in the conference proceedings of the 18th International Congress of Phonetic Sciences in 2015. I would like to thank Patrice Beddor and Andries Coetzee for advice and discussion at every stage of this research. I am also deeply grateful to Holger Mitterer and anonymous reviewers for their helpful comments; Kuniko Nielsen, Julie Boland, Ioana Chitoran, and members of the University of Michigan Phonetics-Phonology Discussion Group for helpful discussions related to this work; audiences at the 18th International Congress of Phonetic Sciences, and the 88th and 89th annual meetings of the Linguistics Society of America, where earlier versions of this work were presented; Jiseung Kim for help with data analyses; and anonymous participants for making this study possible.

Competing Interests

The author has no competing interests to declare.

References

- Allen, J. S., & Miller, J. L. (1999). Effects of syllable-initial voicing and speaking rate on the temporal characteristics of monosyllabic words. *Journal of the Acoustical Society of America*, *106*, 2031–2039. DOI: <https://doi.org/10.1121/1.427949>
- Babel, M. (2012). Evidence for phonetic and social selectivity in spontaneous phonetic imitation. *Journal of Phonetics*, *29*, 155–190. DOI: <https://doi.org/10.1016/j.wocn.2011.09.001>
- Babel, M., & Bulatov, D. (2012). The role of fundamental frequency in phonetic accommodation. *Language & Speech*, *55*, 231–248. DOI: <https://doi.org/10.1177/0023830911417695>
- Babel, M., McGuire, G., Walters, S., & Nicholls, A. (2014). Novelty and social preference in phonetic accommodation. *Laboratory Phonology*, *5*, 123–150. DOI: <https://doi.org/10.1515/lp-2014-0006>
- Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*, 255–278. DOI: <https://doi.org/10.1016/j.jml.2012.11.001>
- Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). lme4: Linear mixed-effects models using Eigen and S4, R package version 1.1-7. <http://CRAN.R-project.org/package=lme4>.
- Boersma, P., & Weenink, D. (2014). *Praat: Doing phonetics by computer* [computer program], version 5.3.84. Retrieved from: <http://www.praat.org/>.
- Cho, T., Jun, S.-A., & Ladefoged, P. (2002). Acoustic and aerodynamic correlates of Korean stops and fricatives. *Journal of Phonetics*, *30*, 198–228. DOI: <https://doi.org/10.1006/jpho.2001.0153>
- Fowler, C. A. (1986). An event approach to the study of speech perception from a direct realist perspective. *Journal of Phonetics*, *14*, 3–28.
- Fowler, C. A. (1996). Listeners do hear sounds, not tongues. *Journal of the Acoustical Society of America*, *99*, 1730–1741. DOI: <https://doi.org/10.1121/1.415237>
- Fowler, C. A., Brown, J. M., Sabadini, L., & Weihing, J. (2003). Rapid access to speech gestures in perception: Evidence from choice and simple response time tasks. *Journal of Memory and Language*, *49*, 396–413. DOI: [https://doi.org/10.1016/S0749-596X\(03\)00072-X](https://doi.org/10.1016/S0749-596X(03)00072-X)
- Gamer, M., Lemon, J., Fellows, I., & Singh, P. (2012). irr: Various Coefficients of Interrater Reliability and Agreement, R package version 0.84. <http://CRAN.R-project.org/package=irr>.

- Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review*, *105*, 251–279. DOI: <https://doi.org/10.1037//0033-295X.105.2.251>
- Goldinger, S. D., & Azuma, T. (2004). Episodic memory reflected in printed word naming. *Psychonomic Bulletin & Review*, *11*, 716–22. DOI: <https://doi.org/10.3758/BF03196625>
- Halle, M., & Stevens, K. N. (1971). A note on laryngeal features. *Quarterly Progress Report*, *101*, 198–212. Research Laboratory of Electronics, MIT.
- Hombert, J.-M., Ohala, J. J., & Ewan, W. G. (1979). Phonetic explanations for the development of tones. *Language*, *55*, 37–58. DOI: <https://doi.org/10.2307/412518>
- Honorof, D. N., Weihing, J., & Fowler, C. A. (2011). Articulatory events are imitated under rapid shadowing. *Journal of Phonetics*, *39*, 18–38. DOI: <https://doi.org/10.1016/j.wocn.2010.10.007>
- Johnson, K. (1997). Speech perception without speaker normalization: An exemplar model. In K. Johnson, & J. W. Mullennix (Eds.), *Talker variability in speech processing* (pp. 145–166). San Diego: Academic Press.
- Johnson, K. (2006). Resonance in an exemplar-based lexicon: The emergence of social identity and phonology. *Journal of Phonetics*, *24*, 485–499. DOI: <https://doi.org/10.1016/j.wocn.2005.08.004>
- Kagaya, R. (1974). A fiberoptic and acoustic study of the Korean stops, affricates, and fricatives. *Journal of Phonetics*, *2*, 161–180.
- Kang, K.-H., & Guion, S. G. (2008). Clear speech production of Korean stops: Changing phonetic targets and enhancement strategies. *Journal of the Acoustical Society of America*, *124*, 3909–3917. DOI: <https://doi.org/10.1121/1.2988292>
- 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. DOI: <https://doi.org/10.1016/j.wocn.2014.03.005>
- Kim, M. (2012). *Phonetic accommodation after auditory exposure to native and nonnative speech* (Doctoral dissertation). Northwestern University. DOI: <https://doi.org/10.1121/1.4755133>
- Kim, M.-R. (2000). *Segmental and tonal interactions in English and Korean: A phonetic and phonological study* (Doctoral dissertation). University of Michigan, Ann Arbor.
- Kim, M.-R., Beddor, P. S., & Horrocks, J. (2002). The contribution of consonantal and vocalic information to the perception of Korean initial stops. *Journal of Phonetics*, *30*, 77–100. DOI: <https://doi.org/10.1006/jpho.2001.0152>
- Knoblauch, K. (2014). psyphy: Functions for analyzing psychophysical data in R, R package version 0.1-9. <https://CRAN.R-project.org/package=psyphy>.
- Kong, E. J., Beckman, M. E., & Edwards, J. (2011). Why are Korean tense stops acquired so early?: The role of acoustic properties. *Journal of Phonetics*, *39*, 196–211. DOI: <https://doi.org/10.1016/j.wocn.2011.02.002>
- Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2014). lmerTest: Tests in Linear Mixed Effects Models, R package version 2.0-20. <http://CRAN.R-project.org/package=lmerTest>.
- Lee, H., & Jongman, A. (2012). Effects of tone on the three-way laryngeal distinction in Korean: An acoustic and aerodynamic comparison of the Seoul and South Kyungsang dialects. *Journal of the International Phonetic Association*, *42*, 145–169. DOI: <https://doi.org/10.1017/S0025100312000035>
- 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*, 117–132. DOI: <https://doi.org/10.1016/j.wocn.2012.12.002>

- Lenth, R. (2018). emmeans: Estimated Marginal Means, aka Least-Squares Means, R package version 1.1. <https://CRAN.R-project.org/package=emmeans>.
- McCrea, C. R., & Morris, R. J. (2005). The effects of fundamental frequency levels on voice onset time in normal adult male speakers. *Journal of Speech, Language, and Hearing Research, 48*, 1013–1024. DOI: [https://doi.org/10.1044/1092-4388\(2005/069\)](https://doi.org/10.1044/1092-4388(2005/069))
- Mitterer, H., & Ernestus, M. (2008). The link between perception and production is phonological and abstract: Evidence from the shadowing task. *Cognition, 109*, 168–173. DOI: <https://doi.org/10.1016/j.cognition.2008.08.002>
- Mitterer, H., & Müsseler, J. (2013). Regional accent variation in the shadowing task: Evidence for a loose perception-action coupling in speech. *Attention, Perception & Psychophysics, 75*, 557–575. DOI: <https://doi.org/10.3758/s13414-012-0407-8>
- Moulines, E., & Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication, 9*, 453–467. DOI: [https://doi.org/10.1016/0167-6393\(90\)90021-Z](https://doi.org/10.1016/0167-6393(90)90021-Z)
- Narayan, C., & Bowden, M. (2013). Pitch affects voice onset time (VOT): A cross-linguistic study. *Proceedings of Meetings on Acoustics, 19*, 060095. DOI: <https://doi.org/10.1121/1.4800681>
- Nielsen, K. (2011). Specificity and abstractness of VOT imitation. *Journal of Phonetics, 39*, 132–142. DOI: <https://doi.org/10.1016/j.wocn.2010.12.007>
- Oh, E. (2011). Effects of speaker gender on voice onset time in Korean stops. *Journal of Phonetics, 39*, 59–67. DOI: <https://doi.org/10.1016/j.wocn.2010.11.002>
- Pardo, J. (2006). On phonetic convergence during conversational interaction. *Journal of the Acoustical Society of America, 119*, 2382–2393. DOI: <https://doi.org/10.1121/1.2178720>
- Pardo, J. (2013). Reconciling diverse findings in studies of phonetic convergence. *Proceedings of Meetings on Acoustics, 19*, 060140. DOI: <https://doi.org/10.1121/1.4798479>
- Pardo, J. S., Gibbons, R., Suppes, A., & Krauss, R. M. (2012). Phonetic convergence in college roommates. *Journal of Phonetics, 40*, 190–97. DOI: <https://doi.org/10.1016/j.wocn.2011.10.001>
- Pardo, J. S., Jordan, K., Mallari, R., Scanlon, C., & Lewandowski, E. (2013). Phonetic convergence in shadowed speech: The relation between acoustic and perceptual measures. *Journal of Memory and Language, 69*, 183–195. DOI: <https://doi.org/10.1016/j.jml.2013.06.002>
- Pierrehumbert, J. (2001). Exemplar dynamics: Word frequency, lenition, and contrast. In J. Bybee, & P. Hopper (Eds.), *Frequency effects and the emergence of lexical structure* (pp. 137–157). Amsterdam: John Benjamins. DOI: <https://doi.org/10.1075/tsl.45.08pie>
- R Development Core Team. (2014). *R: A language and environment for statistical computing* [computer program]. Retrieved from: <http://www.R-project.org/>.
- Schmid, H.-J., & Günther, F. (2016). Towards a unified socio-cognitive framework for salience in language. *Frontiers in Psychology, 7*, 1110. DOI: <https://doi.org/10.3389/fpsyg.2016.01110>
- Shockley, K., Sabadini, L., & Fowler, C. A. (2004). Imitation in shadowing words. *Perception and Psychophysics, 66*, 422–429. DOI: <https://doi.org/10.3758/BF03194890>
- Silva, D. J. (2006). Acoustic evidence for the emergence of tonal contrast in contemporary Korean. *Phonology, 23*, 287–308. DOI: <https://doi.org/10.1017/S0952675706000911>
- The National Institute of the Korean Language. (2005). *A speech corpus of reading-style standard Korean* [DVDs]. Seoul, Korea: The National Institute of the Korean Language.

- Tilsen, S. (2009). Subphonemic and cross-phonemic priming in vowel shadowing: Evidence for the involvement of exemplars in production. *Journal of Phonetics*, 37, 276–296. DOI: <https://doi.org/10.1016/j.wocn.2009.03.004>
- Walker, A., & Campbell-Kibler, K. (2015). Repeat what after whom? Exploring variable selectivity in a cross-dialectal shadowing task. *Frontiers in Psychology*, 6, 546. DOI: <https://doi.org/10.3389/fpsyg.2015.00546>
- Zellou, G., Scarborough, R., & Nielsen, K. (2016). Phonetic imitation of coarticulatory vowel nasalization. *Journal of the Acoustical Society of America*, 140, 3560–3575. DOI: <https://doi.org/10.1121/1.4966232>

How to cite this article: Kwon, H. 2019 The role of native phonology in spontaneous imitation: Evidence from Seoul Korean. *Laboratory Phonology: Journal of the Association for Laboratory Phonology* 10(1):10, pp.1–24. DOI: <https://doi.org/10.5334/labphon.83>

Submitted: 16 February 2017 **Accepted:** 08 April 2019 **Published:** 14 June 2019

Copyright: © 2019 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See <http://creativecommons.org/licenses/by/4.0/>.

 *Laboratory Phonology: Journal of the Association for Laboratory Phonology* is a peer-reviewed open access journal published by Ubiquity Press.

OPEN ACCESS 