SPONTANEOUS SPEECH IMITATION AND CUE PRIMACY

Harim Kwon

University of Michigan
harim@umich.edu

ABSTRACT

This study investigates the influence of cue primacy on spontaneous speech imitation by speakers of Seoul Korean. In Seoul Korean, at least two distinct acoustic cues, stop VOT and *f0* of the post-stop vowel, differentiate aspirated stops from other phonation types. Previous studies have shown that English speakers imitate extended VOTs of voiceless stops both in immediate shadowing [12] and delayed imitation [11]. In this study, Seoul Korean speakers heard Korean aspirated /t^h/ with either extended VOT or raised post-stop *f0*. The realization of these properties in their own /t^h/, /t/, and /t*/ productions were compared before, during, and after exposure. After hearing /t^h/s with extended VOT, participants produced these stops with higher post-stop *f0* as well as longer VOT. After hearing /t^h/s with raised post-stop *f0*, however, the same participants did not lengthen VOT but raised post-stop *f0*. This asymmetry is explained in terms of cue primacy in Seoul Korean.

Keywords: Spontaneous speech imitation, voice onset time, aspiration, cue primacy, Seoul Korean

1. INTRODUCTION

Korean has a three-way laryngeal contrast for voiceless stops: tense, lax, and aspirated. To maintain the full three-way contrast, contemporary Seoul Korean requires at least two distinct acoustic cues, stop VOT and *f0* of the following vowel. For phonological aspiration, post-stop high *f0* has become the primary cue both in perception [6] and production [5]; the VOT difference between aspirated and lax stops is merged for most Seoul Korean speakers although VOT is still relevant for aspirated-tense contrast. This study investigates how cue primacy influences speech imitation, asking whether primary and non-primary cues for aspiration exhibit different imitation patterns.

Previous findings for spontaneous speech imitation show that English speakers imitate artificially extended VOTs of voiceless stops both in shadowing (i.e., immediate repetition of target words) [12], and in delayed imitation [11]. Since VOT is a primary cue for English voiceless stops, it still remains unclear (1) whether extended VOT facilitates imitation when it is not the primary cue, as in Seoul Korean, and (2) whether imitation is strictly tied to the manipulated phonetic property (e.g., long VOT) or it is rather phonological in that other typically co-occurring phonetic properties are also enhanced in imitated productions.

This study examines spontaneous speech imitation in the presence of multiple phonetic cues. Seoul Korean speakers heard and shadowed target speech in which the phonetic information for phonological aspiration was manipulated. The target aspirated stops in this study had either extended VOT or raised post-stop *f0*. By examining how these cues operate in spontaneous imitation, this study aims to reveal whether the target of speech imitation is detailed phonetic properties such as long VOT or high pitch or phonological aspiration. In addition, this study also investigates how lax and tense stops change as a consequence of hearing the target speech with manipulated aspirated stops.

2. METHODS

2.1. Participants

Participants were nineteen native speakers of Seoul Korean (12 females and 7 males), living in Ann Arbor, Michigan. Their ages range from 20 to 31 (mean = 24.4, s.d. = 3.3). No participants reported any history of speech or hearing impairments. Participants were paid for their time.

2.2. Stimuli

Korean words with initial aspirated /t^h/, lax /t/, or tense /t*/ were selected as test words from the NIKL corpus of modern Korean [13]. In addition, words with initial sonorants were selected as fillers. All selected 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 [13]). The word familiarity score was obtained from ten native speakers of Korean who were not participants in the main study. They were asked to rate the familiarity of the target words presented with fillers on a 7-point scale, and only words that obtained a familiarity score of higher than 6.0 on average were selected.

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, comprising half of the /t^h/-initial words and half of the fillers from the reading list.

A male native speaker of Seoul Korean (age = 25) served as a model speaker and recorded the 50 words in the shadowing list. The mean VOT for the model speaker's initial /t^h/s was 58.4 ms (s.d. = 8.6) and the mean *f0* at the midpoint of post-aspirated-stop vowels was 153.6 Hz (s.d. = 4.5).

The model speech was manipulated in two different ways. First, in order to create the high *f0* stimuli, the first pitch period of the vowel that followed word-initial /t^h/ was raised by 20% (in Hz value), and *f0* of the rest of the first vowel was raised proportionately. After manipulation, mean *f0* at midpoint of post-stop vowels was 176.2 Hz (s.d. = 7.1). Second, the long VOT stimuli were created by extending the VOT of word-initial /t^h/ by 60 ms. After manipulation, the mean VOT for /t^h/-initial targets was 119.8 ms (s.d. = 8.1). All manipulations were done using Praat [3].

2.3. Procedure and task

Each participant was tested in two different experimental sessions that were at least two weeks apart from each other. Each experimental session involved target stimuli with a different manipulation, one with raised *f0* and the other with extended VOT. The order of the two experimental sessions was counterbalanced between participants to prevent any potential confounding effect of the testing order.

Each experimental session consisted of warm-up, baseline, shadowing, and test blocks, using a slightly modified version of the word-naming imitation paradigm [1, 4, 11]. Participants were seated in front of a laptop in a sound-attenuated booth in the Phonetics Laboratory at the University of Michigan. 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 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 *f0* or VOT manipulation were played via headphones 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. The test block was conducted in the same way as the baseline.

2.4. Measurements

Stop VOT and post-stop *f0* of each token in the participants' baseline, shadowing, and test productions were measured. VOTs for word-initial stops were taken from the release burst to the beginning of true modal voicing of the post-stop vowel, as seen in spectrograms in F2 and above. Thus, VOT in this study includes any breathy voiced portion of the post-stop vowel. Post-stop *f0* measures were taken at the temporal midpoint of the first vowel of each word.

2.5. Statistical analyses

All data were analyzed using R with the lme4 package [2]. Different manipulation types (High *f0* vs. Long VOT) were separately analyzed. Also separately analyzed were changes in aspirated /t^h/s in all different Production types (/t^h/-Only models) and changes in all Stop types (Aspirated /t^h/, Lax /t/, and Tense /t*/) in baseline and test productions (3-Stop models). Fixed effects for the /t^h/-Only models were Production types (Base, Sh1, Sh2, Sh3, and Test), Presence of exposure (Shadowed vs. Unheard), and speaker Gender (without interaction terms). The 3-Stop models included Stop types (Aspirated /t^h/, Lax /t/, and Tense /t*/) and its interaction with Production types (Base vs. Test), as well as Presence of exposure and Gender, as fixed effects.

For each analysis, separate linear mixed effects models were constructed with stop VOT and post-stop *f0* as the dependent variables. For VOT models, the rest of word duration (Rest = total word duration – VOT) was included as a predictor to make sure that the changes in VOT were not due to global changes in speech rate. Similarly, to make sure that the observed changes in post-stop *f0* were not due to a global shift in pitch range, filler words were included in the *f0* models by including the interaction between Word types (Aspirated /t^h/-words vs. Fillers) and Production types. Intercepts for speakers and words were included as random effects, as well as by-speaker random slopes. Parameter-specific *p*-values were obtained using the Satterthwaite approximation implemented in the lmerTest package [7].

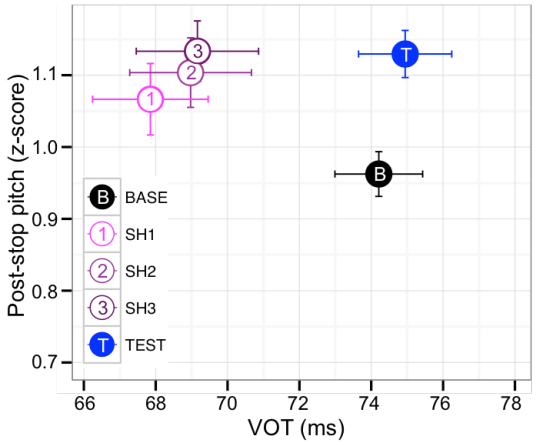
3. RESULTS

3.1. Effects of enhanced primary cue (high *f0*)

3.1.1. Aspirated /t^h/ in base-shadowing-test blocks

Fig. 1 presents average VOT and *f0* of aspirated /t^h/s in the different Production types in the high *f0* condition. For clarity, post-stop *f0* is presented in *z*-scores in all figures, although statistical analyses were performed on raw Hz values. Throughout the paper, error bars represent 95% confidence intervals.

Figure 1: Changes in aspirated /t^h/ in high pitch condition



In the high *f0* condition, Production types had a significant shortening effect on VOT during each shadowing repetition relative to Base [Sh1: β=−6.43, t=−5.07, p<0.001; Sh2: β=−5.27, t=−4.23, p<0.001; Sh3: β=−5.14, t=−3.33, p=0.003]. The VOT difference between Base and Test was not significant.

The effects of Production types on post-stop *f0* were significant but only the pairwise comparison between Base-Test was found to be significant [β=5.10, t=3.32, p<0.001]. None of the three repetitions of shadowing productions was different from Base [|t|<2, p>0.1]. However, examining the coefficients of the *f0* model by speaker revealed that there were three outlier speakers. These outlier speakers were all females, who apparently imitated the male model speaker by lowering their pitch. Excluding these three outliers, the same linear mixed effects model was fitted again, and the post-stop *f0* was significantly higher in both Sh2 and Sh3 than Base [|t|>3, p<0.005]. The *f0* difference between Sh1 and Base was marginally significant [t=1.91, p=0.08], and the post-stop *f0* was significantly higher in Test than Base [β=6.16, t=3.75, p=0.002]. The change in post-stop *f0* was not due to a global pitch range shift: post-sonorant *f0* in filler words showed a significant decrease between Base and Test [β=−4.54, t=−3.14, p=0.006], and no significant changes between Base and Sh1, 2, 3.

A separate linear mixed effects model was fitted to the outlier speakers only, which revealed that their post-stop *f0* in shadowing productions was indeed lower than their baseline, especially in the earlier repetitions of shadowing [Sh1: β=−14.68, t=−3.40, p=0.04; Sh2: β=−16.43, t=−4.05, p=0.02; Sh3: β=−12.48, t=−2.55, p=0.08]. These outlier speakers' post-stop *f0* in Test was not different from their Base [β=−0.47, t=−0.18, p=0.87].

3.1.2. Stops of different phonation types in base-test

The effects of Production types (Base vs. Test) on stop VOT were not significant for any of the Stop types under investigation. That is, participants did not adjust VOTs for their /t^h/, /t/ or /t*/ productions after hearing and shadowing /t^h/s with artificially raised post-stop *f0*.

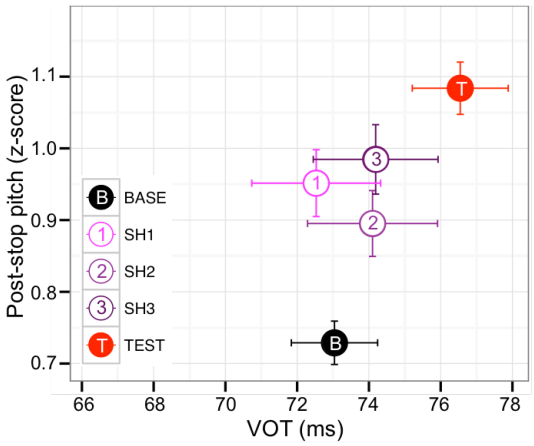
On the other hand, Production types had significant effects on post-stop *f0* for both aspirated /t^h/ and lax /t/. Post-/t^h/ *f0* increased in Test [β=5.11, t=3.34, p=0.003], while post-/t/ *f0* decreased [β=−3.24, t=−2.59, p=0.017]. Post-/t*/ *f0* did not change significantly.

3.2. Effects of enhanced non-primary cue (long VOT)

3.2.1. Aspirated /t^h/ in base-shadowing-test blocks

Fig. 2 presents average VOT and *f0* of participants' aspirated /t^h/ productions before and after hearing /t^h/-initial words with lengthened VOT.

Figure 2: Changes in aspirated /t^h/ in long VOT condition



In the long VOT condition, the effects of Production types on stop VOT were significant only between Base and Test [β=2.99, t=2.16, p=0.035]. Exposure to aspirated /t^h/s with extended VOT did not induce significant imitation effects in stop VOT during shadowing productions [|t|<0.5, p>0.1].

On the other hand, post-stop *f0*s in Sh1 and Sh3 were significantly higher than those in Base [Sh1: β=4.99, t= 2.11, p=0.043; Sh3: β=4.93, t=2.01, p=0.045]. Post-stop *f0* in Test also showed a significant increase from Base [β=9.33, t=4.57, p<0.001], while post-sonorant *f0* in filler words significantly decreased [β=−7.31, t=−3.76, p<0.001]. Post-stop *f0* in Sh2 was not significantly different from the baseline counterpart [t=1.35, p=0.18].

3.2.2. Stops of different phonation types in base-test

Among the three phonation types examined, the effects of Production types on stop VOT were significant only for aspirated stops (i.e., only for the stop category whose phonetic properties were manipulated during shadowing): Test /t^h/ had longer VOT than Base /t^h/ [β=3.04, t=2.20, p=0.04]. Stop VOT of lax and tense stops was not different between Base and Test.

On the other hand, the effects of Production types on post-stop *f0* were highly significant for all Stop types: after hearing and shadowing aspirated /t^h/s with artificially lengthened VOT, *f0* after aspirated stops had a significant increase [β=9.32, t=4.69, p<0.001], while post-lax-stop and post-tense-stop *f0*s showed significant decreases [β=−5.86, t=−3.09, p=0.006; β=−2.83, t=−3.09, p=0.004, respectively].

4. DISCUSSION AND CONCLUSION

The current findings demonstrate a clear asymmetry between primary and non-primary cues in spontaneous speech imitation of Seoul Korean aspirated stops. Although both primary and non-primary cues facilitated imitative effects, the details of the imitation patterns were distinct. An enhanced non-primary cue for stop aspiration (long VOT) induced increases in both primary and non-primary cues. That is, the Seoul Korean speakers "imitated" exaggerated stop aspiration cued by long VOT not only by lengthening their own VOT for aspirated stops but also by raising their *f0* after those stops. However, an enhanced primary cue (high *f0*) did not have similar effects on the non-primary cue. After hearing aspirated stops with raised post-stop *f0*, the same participants only imitated the manipulated property; they did not lengthen VOT.

These results indicate that, in speech imitation, exposure to an enhanced phonetic property can influence production not only of that property but also of other phonetic properties if they are important to the targeted phonological category. The participants in this study appears to have adjusted the cue that they would primarily use to enhance stop aspiration (i.e., post-stop *f0*), in addition to merely imitating the cue that the stimuli they heard employed to enhance stop aspiration.

What caused the decrease in stop VOT during the shadowing blocks of the high *f0* condition, accompanying an increase in post-stop *f0*? One possibility is that participants imitated the model speaker's relatively short VOT (about 60 ms). However, the same participants did not imitate extended VOT (about 120 ms) in their shadowing productions of the long VOT condition, and it seems unlikely that they imitated a small difference in one condition but not a much larger difference in another condition. This reverse relation between stop VOT and post-stop *f0*, in fact, has been reported in several previous studies [8, 10]. Contrary to the pitch perturbation caused by stop (de-)voicing, high pitch causes the vocal cords to be stiff, reducing the time for the vocal cords to adduct and start voicing.

Consistent with previous findings that phonetic imitation is not limited to specific words, but generalizes to novel words sharing the same phoneme [11], the present study did not find any effects of Presence of exposure (Shadowed vs. Unheard). Words beginning with initial aspirated stops showed the same imitation effects regardless of whether the specific word was present or not during the shadowing block.

However, examining the imitation effects on stops of other phonation types suggests that speech imitation may not be limited to a specific phonological category or feature. In the test production blocks, participants not only raised their *f0* following aspirated stops but also lowered *f0* after lax stops (in both conditions) and tense stops (in the long VOT condition only). It remains unclear why post-tense-stop *f0* decreased only in the long VOT condition. Nevertheless, it seems the participants "imitated" the enhanced aspiration by maximizing the difference in post-stop *f0* between aspirated stops and other stops, as post-stop *f0* is the primary cue for stop aspiration.

In sum, this study suggests that the target of speech imitation is not an isolated acoustic parameter such as long VOT or high *f0*, but rather complex phonological categories such as stop aspiration. These findings further extend the claim that speech imitation is phonological [9]. A phonetic property that has a primary contrastive role (here, high *f0*) appears to have precedence over a less important property in speech imitation.

5. REFERENCES

[1] Babel, M. 2012. Evidence for phonetic and social selectivity in spontaneous phonetic imitation. *Journal of Phonetics* 29, 155-190.

[2] 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>.

[3] Boersma, P., Weenink, D. 2014. Praat: Doing phonetics by computer (version 5.3.84). [Computer program], <http://www.praat.org/>.

[4] Goldinger, S. D. 1998. Echoes of echoes? An episodic theory of lexical access. *Psychological Review* 105, 251-279.

[5] 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.

[6] 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.

[7] 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>.

[8] 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.

[9] Mitterer, H., Ernestus, M. 2008. The link between speech perception and production is phonological and abstract: Evidence from the shadowing task. *Cognition* 109, 168-173.

[10] Narayan, C., Bowden. M. 2013. Pitch affects voice onset time (VOT): A cross-linguistic study. *Proceedings of Meetings on Acoustics* 19.

[11] Nielsen, K. 2011. Specificity and abstractness of VOT imitation. *Journal of Phonetics* 39, 132-142.

[12] Shockley, K., Sabadini, L., Fowler, C. A. 2004. Imitation in shadowing words. *Perception and Psychophysics* 66, 422-429.

[13] 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.