GENERALIZATION OF DIMENSION-BASED STATISTICAL LEARNING IN WORD RECOGNITION

Lori L. Holt[a] *& Kaori Idemaru*[b]

[a]Carnegie Mellon University, USA; [b]University of Oregon, USA
holt@andrew.cmu.edu; idemaru@uoregon.edu

ABSTRACT

Recent research demonstrates that the diagnosticity of an acoustic dimension for speech categorization is relative to its relationship to the evolving distribution of dimensional regularity across time, and not simply to its fixed value along the dimension. Two studies examine the nature of this *dimension-based statistical learning* in online word recognition, testing generalization of learning across talkers and across phonetic categories. The results indicate that dimension-based statistical learning generalizes across talkers, but is specific to experienced phonetic categories.

Keywords: speech perception, statistical learning, dimension-based learning, cue weighting

1. INTRODUCTION

The long-term regularities of speech input may not adequately capture the regularity of speech in the short-term, such as in the case of non-native accented speech. Speech perception must remain flexible enough to adapt to such variability. Native English speakers learning Korean, for instance, use the canonical English relationship of voice onset time (VOT) and fundamental frequency (F0; higher F0 for voiceless consonants, [7]) when producing Korean consonants, even though this relationship is not characteristic of Korean [6]. This non-native instantiation of Korean thus violates typical correlations among dimensions defining native-Korean speech categories and presents a perceptual challenge for native Korean listeners.

Recent research [5] found that online speech processing rapidly adjusts the perceptual weight of acoustic dimensions in response to perturbations of correlations between acoustic dimensions like VOT and F0. In these experiments, listeners heard artificially "accented" utterances of rhymes *beer*, *pier*, *deer* or *tear,* in which the correlation between F0 and voicing categories was reversed from the English norm [1] such that higher F0s were paired with voiced sounds (*beer* and *deer*) and lower F0s

were paired with voiceless sounds (*pier* and *tear*). F0 typically influences voicing judgments when VOT is perceptually ambiguous [4]. However, listeners down-weighted reliance on the F0 dimension in word recognition within just a few trials of experience with the reversed F0/VOT correlation such that it no longer influenced voicing categorization. These results demonstrate rapid acoustic *dimension-based statistical learning;* listeners track relationships between acoustic dimensions in online speech processing and the diagnosticity of an acoustic dimension for a phonetic category is not simply a function of its value along the acoustic dimension. Rather, it is evaluated relative to evolving regularities between acoustic dimensions in short-term experience. This perceptual tuning process is likely to be important for understanding how listeners deal with the acoustic perturbations to speech resulting from accent, dialect and dysarthria. The present studies investigate whether this learning is talker- (Exp.1) and/or phoneme-specific (Exp. 2).

2. EXPERIMENT 1

2.1. Methods

Twenty-eight native-English listeners with normal hearing participated.

2.1.1. Stimulus creation

Natural utterances of *beer, pier, deer* and *tear* ([bɪər], [pɪər], [dɪər], and [tɪər]) were digitally recorded (22.05 kHz) from utterances of a female monolingual native speaker of English (LLH, Voice 1). Words were spoken in isolation in citation form. Using these utterances as endpoints, VOT was manipulated in seven 10-ms steps from -20 ms to 40 ms for the *beer/pier* series and -10 ms to 50 ms for the *deer/tear* series (pilot categorization tests indicated category boundaries at about 10-ms VOT for *beer/pier* series and 20-ms VOT for *deer/tear*). An instance of *pier* and an instance of *tear* were chosen based on clarity and their roughly equivalent durations. The waveforms

of these voiceless endpoints were edited (see [3]) to create a *beer* to *pier* and a *deer* to *tear* series. The first 10 ms of the original voiceless productions were left intact to preserve the consonant bursts. Manipulation of VOT across the series was accomplished by removing approximately 10-ms segments (with minor variability so that edits were made at zero-crossings) from the waveform using Praat 5.0 [2]. For the negative VOT values, pre-voicing was taken from voiced productions of the same speaker and inserted before the burst in durations varying from -20 to 0 ms in 10 ms steps.

The two series were manipulated such that the F0 onset frequency of the vowel, [ɪ], following the word-initial stop consonant was adjusted from 220 Hz to 300 Hz across nine 10-Hz steps. For each stimulus, the F0 contour of the original production was measured and manually manipulated using Praat 5.0 to adjust the target onset F0 values. The F0 remained at the target frequency for the first 80 ms of the vowel; from there, it linearly decreased over 150 ms to 180 Hz.

A second set of test stimuli to investigate generalization was created by applying Praat's change-gender function to the female Voice 1 stimuli. This function manipulates F0 and formant frequencies independently [8]. By decreasing the formant frequencies of the original test stimuli by a factor of 0.8 while holding F0 at the original values, we created a "male" version (Voice 2) of the test stimuli that 71% of our 38 participants (N=27) found to be a convincing male/female talker difference; only these participants were included in data analyses.

2.1.2. Procedure

In Block 1, native English listeners heard speech with the Canonical English F0/VOT correlation: voiced stops had lower F0s whereas voiceless stops had higher F0s in the following vowel (Figure 1). In Block 2, listeners heard speech with the F0/VOT correlation Reversed: voiced stops were associated with higher F0s and voiceless stops with lower F0s. In a block, the exposure stimuli (open symbols) were presented 15 times each in random order. All exposure stimuli were produced in Voice 1. The VOT-neutral test stimuli (filled symbols) spoken in Voice 1 and Voice 2 were each presented 10 times per block, interspersed randomly among the exposure stimuli. Trials proceeded continuously across the two

blocks as listeners performed a four-alternative word-recognition task. The block structure was implicit. Participants were not informed that the experiment was divided into separate blocks, that the nature of the acoustic cues would vary, or that they would hear words spoken in different voices. The entire session was completed in approximately 45 minutes.

Figure 1: Schematic illustration of stimulus sampling in Block 1 (left, Canonical English F0/VOT correlation) and Block 2 (right, Reversed F0/VOT correlation) as a function of VOT and F0.
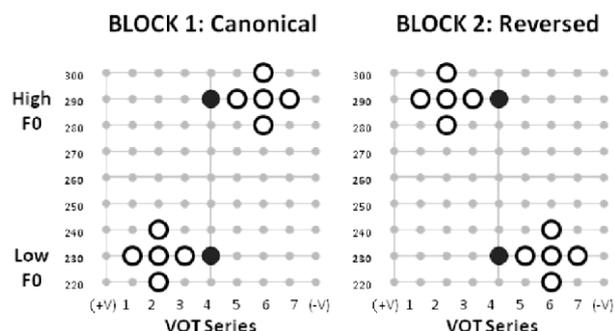
Following the word-recognition task, participants categorized 10 random presentations each of the 4 test stimuli (2 F0 levels x 2 Voices) as a "male" voice or "female" voice. Eleven listeners who failed to accurately categorize the two voices were excluded from analyses on the conservative logic that generalization cannot be assessed adequately among listeners who did not reliably distinguish the voices. Three listeners who experienced technical problems during the word-recognition task were also excluded from the analyses. Data are reported for 24 listeners.

2.2.   Results

2.2.1. Responses to exposure stimuli

Listeners' responses to exposure stimuli with unambiguous voiceless and voiced VOT values verify that participants used the primary cue, VOT, in word recognition (-10 ms for [b] and 30 ms for [p]) [voiced, M =93.1, SE = .63; voiceless, M = 94.6, SE = .51].

2.2.2. Responses to test stimuli

Figure 2 shows the results. Percent voiceless responses were submitted to a 2 (Block: Canonical vs. Reversed) x 2 (F0: High vs. Low) x 2 (Voice: Exposed vs. New) repeated-measures ANOVA separately for *beer/pier* and *deer/tear*.

Figure 2: Percent voiceless responses to the Exposed Voice (top row) and the New Voice (bottom row) as a function of High and Low F0 in blocks with a Canonical (high, voiceless) or Reversed (high, voiced) FO/VOT correlation.
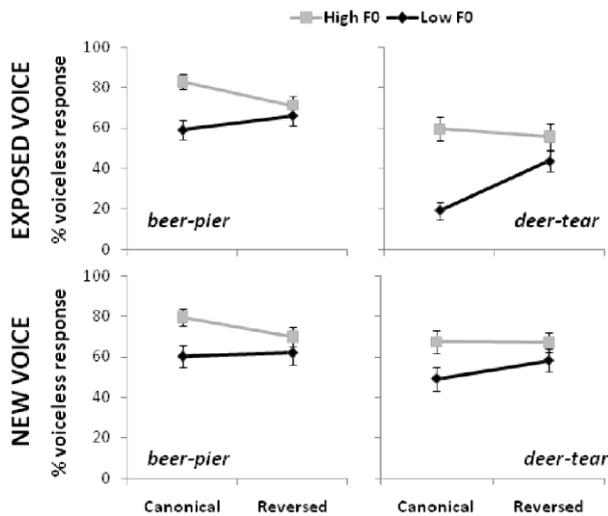
For *beer/pier*, there was a significant main effect of F0 [$F(1, 23) = 17.965$, $p < .001$] and a significant F0 x Block interaction [F0 x Block, $F(1, 23) = 7.987$, $p < .01$]. There was no main effect of Voice, or any interactions involving Voice; thus, the percent voiceless scores were collapsed across Voice. Post-hoc paired t-tests reveal an effect of F0 in the Canonical Block [$t(23) = -4.998$, $p < .0001$] but not in the Reversed Block [$t(23) = -1.551$, $p = .1346$ (alpha = .025)]. Listeners exhibited the expected influence of F0 on voicing in Block 1, but exposure to the Voice 1's speech with a Reversed F0/VOT correlation in Block 2 led listeners to down-weight F0 as a cue to voicing. This pattern was generalized across talker, to Voice 2.

For *deer/tear,* the ANOVA revealed significant main effects of F0, Block, and Voice as well as significant Voice x F0 and Block x F0 interactions [Voice, $F(1, 23) = 26.154$, $p < .0001$; Block, $F(1, 23) = 6.047$, $p < .05$; F0, $F(1, 23) = 40.419$, $p < .0001$; Voice x F0, $F(1, 23) = 4.607$, $p < .05$; Block x F0, $F(1, 23) = 25.439$, $p < .0001$]. The Voice x F0 interaction indicates that F0 effect was different depending on the voice regardless of blocks. The F0 effects were 26.3 and 13.5 for Voice 1 (exposed) and Voice 2 (new), both statistically significant [Voice 1, $t(23) = -5.836$, $p < .0001$; Voice 2, $t(23) = .0049$ (alpha = .025)]. Listeners used F0 as a cue to voicing to a greater degree in categorizing words produced in the exposed voice than words produced in the new

voice. More importantly, the Block x F0 interaction indicates that F0 effect was different across Blocks regardless of Voice. The F0 effects were 29.4 and 10.4 in Block 1 and 2, each of which was statistically significant [Block 1, $t(23) = -8.153$, $p < .0001$; Block 2, $t(23) = -2.817$, $p = .0098$ (alpha = .025)]. These results indicate that although the F0 effect did not disappear in the Reverse correlation block, it was greatly diminished. Moreover, the modulation of the F0 effect by Block was consistent across the two voices, indicating that learning in one voice generalized to another voice.

3. EXPERIMENT 2

In this experiment, we investigate whether learning generalizes to voicing categories with which listeners have not had experience with an F0/VOT reversal.

3.1. Methods

Fourteen normal-hearing, native-English listeners participated.

3.1.1. Procedure

In Block 1, listeners heard speech with the familiar, canonical English F0/VOT correlation only for the *beer-pier* series: the vowel in *beer* had lower F0s whereas the vowel in *pier* had higher F0s (Canonical block). In Block 2, listeners heard *beer* and *pier* with the reversed F0/VOT correlation such that utterances of *beer* and *pier* had an F0/VOT correlation opposite their long-term experience with English (Reversed block). Thus, across blocks the correlation of F0 and VOT in the *beer-pier* series shifted from the natural pattern that characterizes the long-term regularities of English to a correlation pattern opposite that of English. In each block, each exposure stimulus was presented 30 times in a random order. To examine the phoneme generalization, test stimuli included VOT-neutral tokens of both *beer-pier* series and *deer-tear* series, whereas exposure stimuli included only *beer-pier* series. These VOT-neutral test stimuli were each presented 20 times per block, interspersed randomly among the exposure stimuli.

There were 600 exposure trials (1 type (*beer-pier*) x 10 exposure sounds x 30 repetitions x 2 blocks) and 160 test trials (2 types (*beer-pier*, *deer-tear*) x 2 F0 levels x 20 repetitions x 2 blocks). Trials proceeded continuously, with a shift in the F0/VOT correlation half way, and

listeners performed the same word-recognition task throughout the experiment. The block structure was not apparent in the nature of the task.

3.2. Results

3.2.1. Responses to exposure stimuli

Listeners' responses to exposure stimuli with unambiguous voiceless and voiced VOT values verify that participants used the primary cue, VOT, in word recognition (-10 ms for [b] and 30 ms for [p]). Mean percentages of expected (correct) responses were high [voiced, M = 92.5, SE = .53; voiceless, M = 94.9, SE = .44].

3.2.2. Responses to test stimuli

Figure 3 reports the mean percent voiceless responses for exposed words (*pier*) and new words (*tear*) across Canonical and Reversed blocks. In categorizing exposed words (*beer* and *pier*), listeners' voiceless responses to high-F0 stimuli (gray line) decreased from the Canonical block to the Reversed block, whereas voiceless responses to low-F0 stimuli (dark line) increased. In categorizing new words (*deer* and *tear*), listeners responses did not show a substantial change.

Figure 3: Percent voiceless responses to Exposed Words (left) and New Words (right) as a function of High and Low F0 in blocks with a Canonical (high, voiceless) or Reversed (high, voiced) F0/VOT correlation.
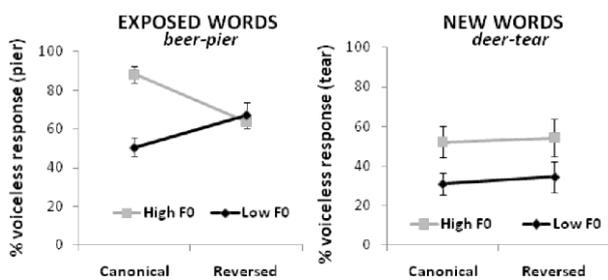
A 2 (Block: Canonical, Reversed) x 2 (F0: high, low) ANOVA was conducted separately for voiceless responses to Exposed words (*pier*) and New words (*tear*). The test for Exposed words returned significant main effects of F0 and a significant Block x F0 interaction [F0, $F(1, 13) = 14.735$, $p < .01$; Block*F0, $F(1, 13) = 56.760$, $p < .0001$]. Post-hoc tests indicated that the F0 effect was significant in the Canonical block [$t(13) = -6.004$, $p < .001$]. Thus, the F0 effect for exposed words was modulated across exposure blocks. Responses to New words patterned differently from responses to the Exposed words. The test for

New words retuned significant main effect only for F0, and no significant Block*F0 interaction [F0, $F(1, 13) = 20.464$, $p < .01$]. This indicates that the F0 effect persisted across the blocks, and the magnitude of the effect was not modulated as a function of the exposure characteristics. Taken together these results indicate that dimension-based learning by *beer-pier* words did not generalize to the categorization of *deer-tear* words. Learning did not generalize to a new phoneme or a new place of stop articulation (from bilabial to alveolar) and did not extend, broadly, to voicing.

4. CONCLUSION

Listeners rely on local input regularities to dynamically "tune" long-term representations by tracking dimensional relationships in online speech processing [5]. Relatively more reliable perceptual sources of information (unambiguous VOT) may adjust perception of less-reliable sources (F0) and perceptual decisions appear to be made using all available information, including prior knowledge. The current findings demonstrate that dimension-based statistical learning is not a talker-contingent process and suggest that learning does not occur at an abstract level like "voicing", but instead is specific to the details of experienced regularities.

5. REFERENCES

[1] Abramson, A.S., Lisker, L. 1985. Relative power of cues: F0 shift versus voice timing. In Fromkin, V.A. (ed.), *Phonetic linguistics: Essays in Honor of Peter Ladefoged.* Orlando: Academic Press, 25-33.

[2] Boersma, P., Weenink, D. 2010. Praat: Doing phonetics by computer [Computer program], Version 5.0. *http://www.praat.org/*

[3] Francis, A.L., Baldwin, K., Nusbaum, H.C. 2000. Effects of training on attention to acoustic cues. *Perception and Psychophysics* 62(8), 1668-1680.

[4] Francis, A.L., Kaganovich, N., Driscoll-Huber, C.J. 2008. Cue-specific effects of categorization training on the relative weighting of acoustic cues to consonant voicing in English. *J. Acoust. Soc. Am.* 124, 1234-1251.

[5] Idemaru, K., Holt, L.L. In press. Word recognition reflects dimension-based statistical learning.

[6] Kim, M.R., Lotto, A.J. 2002. An investigation of acoustic characteristics of Korean stops produced by non-heritage learners. *The Korean Language in America*, 7, 177-188.

[7] Kohler, K.J. 1982. F0 in the production of lenis and fortis plosives. *Phonetica 39*(4-5), 199.

[8] Moulines, E., Charpentier, F. 1990. Pitch-synchronous processing techniques for text-to-speech synthesis using diphones. *Speech Communication* 9, 453-467.