Examining Talker and Phoneme Generalization of Dimension-based Statistical Learning in Speech Perception

Kaori Idemaru[1] *& Lori L. Holt*[2]

[1]University of Oregon, USA
[2]Carnegie Mellon University, USA

idemaru@uoregon.edu;loriholt@cmu.edu

Abstract

Speech perception flexibly adapts to short-term regularities of the ambient speech input. Recent research demonstrates that the function of an acoustic dimension for speech categorization at a given time 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 is specific to the experienced regularities, resisting transfer across talkers or phonetic categories.

Index Terms: speech perception, statistical learning, dimension-based learning, cue weighting

1. Introduction

Speech processing exhibits a dual nature. On the one hand listeners possess sensitivity to long-term regularities of the native language; on the other, they flexibly adapt and retune perception to adjust to short-term deviations arising from the idiosyncrasies of individual speakers in a manner that is helpful in accommodating acoustic variability.

Recent research found that online speech processing rapidly adjusts the perceptual weight of acoustic dimensions defining speech categories in response to perturbations of long-term regularities [3]. In these experiments, listeners heard artificially "accented" rhymes *beer*, *pier*, *deer* or *tear*, in which the correlation between the F0 of the vowel onset and voicing categories was reversed from the English norm: higher F0s were paired with voiced stops (*beer* and *deer*) and lower F0s were paired with voiceless stops (*pier* and *tear*). Within just a few trials of experience with this reversed F0/VOT correlation, listeners down-weighted reliance on F0 such that it no longer influenced 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 fixed function of its value along the acoustic dimension. Rather, it is evaluated relative to evolving regularities between acoustic dimensions in the input in short-term experience. This perceptual tuning is likely to be important for understanding how listeners deal with the acoustic perturbations to speech resulting from accent, dialect and dysarthria.

The current study investigates generalization of this learning. In a prior study, we observed generalization across talker [4], but this was a relatively weak test because the generalization voice was resynthesized from the exposure voice and, though heard as a distinct talker, may have shared critical acoustics with the exposure voice. Here, we test whether learning generalizes to an entirely new voice. In contrast to talker generalization, learning did not generalize across place-of-articulation in the previous research [4]. The current study investigates the implications of this result by examining whether listeners are capable of tracking simultaneously, opposing correlations for *beer-pier* and *deer-tear* stimuli, or whether they aggregate statistical information across the voicing contrasts.

2. Experiment 1

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

2.1. Methods

Fifteen native-English listeners with normal hearing participated in the word recognition task.

2.1.1. Stimuli

The stimuli, *beer, pier, deer* and *tear* ([bɪər], [pɪər], [dɪər], and [tɪər], used in the previously-published work [3] served as stimuli. The stimuli were created based on natural utterances of *pier* and *tear* produced in isolation by a female monolingual native speaker of mid-west American English (second author, Talker 1). 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*). 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 [5]. The first 10 ms of the original voiceless productions were left intact to preserve the consonant bursts. 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 was created based on natural utterances of *pier* and *tear* produced by a male monolingual native speaker of mid-west American English (Talker 2) to investigate talker generalization. An instance of *pier* and an instance of *tear* were chosen based on their roughly equivalent durations with the Talker 1 stimuli. Using the methods described above, F0 and VOT were manipulated such that Talker 1 and Talker 2 were equivalent in terms of F0 and VOT. This manipulation allowed us to control for F0 for the two voices while retaining lower formant frequencies of Talker 2.

2.1.2. Procedure

A categorization task examined the baseline effect of F0. In this task, listeners categorized 4 series of stimuli: *beer/pier* and *deer/tear* continua in Talker 1 and 2 voices varying along VOT (9 steps: -20, -10, 0, 5, 10, 15, 20, 30, 40 ms for *beer-pier*, -10, 0, 10, 15, 20, 25, 30, 40, 50 ms for *deer-tear*) and F0 (2 levels: 230 and 290 Hz). Each stimulus was presented 5 times, except for the boundary (ambiguous) VOT stimuli (step 5), which were presented 10 times so that the number of presentation of these critical stimuli is consistent with subsequent tests. A total of 400 trials were presented in the baseline test, blocked for *beer/pier* and *deer/tear* types and voices, with blocks counter-balanced across participants.

Participants were seated in front of a computer monitor in a sound booth. Each trial presented a spoken word diotically over headphones (Beyer DT-150) and visual icons corresponding to the two response choices (clip-art pictures of a beer and a pier, or a deer and a tear), each with a designated key number, were presented on a monitor. The experiment was delivered under the control of E-prime experiment software (Psychology Software Tools, Inc.). Participants were instructed to press a key corresponding to the picture of the word they heard as quickly as possible.

The word recognition task, which immediately followed the baseline test, exposed listeners to a shifting F0/VOT correlation and monitored their reliance on F0 through the course of the task. In Block 1, 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 Block 3, the correlation returned to Canonical English. The 20 exposure stimuli (open symbols) were presented 10 times per block in random order. All exposure stimuli were produced by Talker 1. The 4 VOT-neutral test stimuli (filled symbols) produced by Talker 1 and Talker 2 were each presented 10 times per block, interspersed randomly among the exposure stimuli.

Trials proceeded continuously across the three blocks as listeners performed the 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.

Following the word-recognition task, participants categorized 5 random presentations each of the 4 test stimuli (2 F0 levels x 2 Voices) as a "male" voice or "female" voice. Three listeners who failed to accurately categorize the two voices above 85% accuracy were excluded from analyses on the conservative logic that generalization cannot be assessed adequately among listeners who did not reliably distinguish the voices.

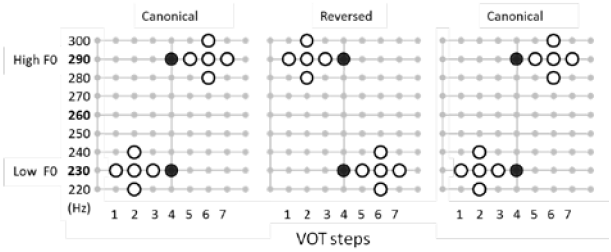

Figure 1: *F0/VOT correlation in stimuli across experimental blocks. Open symbols are exposure stimuli and filled symbols are test stimuli.*

2.2. Results

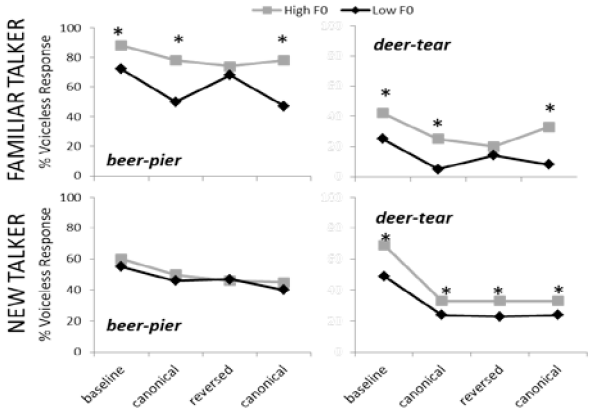

Figure 2: *Percent voiceless responses across four blocks for familiar talker (Talker 1) and new talker (Talker 2) stimuli. A star indicates statistical significance.*

A 9 x 2 (VOT x F0) ANOVA was run on the mean percent voiceless responses to baseline categorization separately for *beer/pier* and *deer/tear* stimuli for each of the two talkers. The statistical report of these tests are abbreviated for the interest of space and as they do not pertain to the central part of the research question. More important is the comparisons between the baseline and subsequent experimental blocks, and those results are reported fully below.

Tests on response to Talker 1 baseline stimuli indicated significant main effects of VOT and F0, and a significant VOT x V0 interaction for both *beer-pier* and *deer-tear* ($p < .001$ for all). Post-hoc paired-sample t-tests indicated that F0 effect (i.e., difference in percent voiceless response between high F0 and low F0) was significant at VOT steps 4 and 5 of VOT for *beer/pier* stimuli and at VOT steps 4, 5, and 6 for *deer/tear* stimuli ($p < .006$, alpha adjusted to .006 for 9 comparisons). Specifically the F0 effect was significant at VOT step 5, which was used as the test VOT value during the subsequent word recognition task.

Tests on response to Talker 2 stimuli indicated a significant main effect of VOT ($p < .001$), and no significant main effect of F0 or significant VOT x F0 interaction for

beer/pier stimuli, whereas there were significant main effects of VOT and F0, and a significant VOT x F0 interaction for *deer/-tear* stimuli ($p < .05$ for all). Post-hoc tests indicated that F0 effect approached significance at VOT step 5 ($p = .042$). These results indicate that listeners showed F0 effects to Talker 1 stimuli (both *beer/pier* and *deer/tear*) but they showed F0 effects only to the *deer/tear* stimuli of Talker 2 at the baseline.

Figure 2 reports the mean percent voiceless response for test stimuli across baseline and three blocks for the familiar talker (Talker 1, top) and the new talker (Talker 2, bottom). Down-weighting of F0 in the reversed block observed in response to the familiar talker was not observed in response to the new talker.

A 4 x 2 (Block x F0) ANOVA was run on percent voiceless response for each of *beer-pier* and *deer-tear* tests for each of familiar talker (Talker 1) and new talker (Talker 2) stimuli. The test for familiar talker's *beer/pier* indicated significant main effects of Block and F0, and a significant F0 x Block interaction [Block: $F(3, 33) = 4.511$, $p = .009$, F0: $F(1, 11) = 45.419$, $p < .001$, F0*Block: $F(3, 33) = 4.199$, $p = .013$]. Here and in the subsequent analyses, baseline was included as one of the four blocks. Paired-sample t-tests indicated that F0 effect was significant in the baseline and canonical blocks ($p < .013$ for all, alpha adjusted to .013 for 4 comparisons), but not in the reversed block. The results were similar for familiar talker's *deer/tear* test. The ANOVA indicated significant main effects of Block and F0, and the F0 x Block interaction approached statistical significance [Block: $F(3, 33) = 5.065$, $p = .005$, F0: $F(1, 11) = 26.349$, $p < .001$, F0*Block: $F(3, 33) = 2.573$, $p = .071$]. Given the marginal F0*Block interaction, the F0 effect was examined in each block. Paired-sample t-tests indicated that the F0 effect was significant in the baseline and canonical blocks, but not in the reversed blocks ($p < .013$ for all, alpha adjusted for 4 comparisons). These results replicated prior findings that listeners modulate the weight that they give to F0 in the perception of voicing contrast according to the input F0/VOT correlation.

On the contrary, the ANOVA for Talker 2's *beer/pier* test indicated no significant main effects or significant interactions. This indicates that lack of an influence of F0 in categorizing Talker 2's *beer/pier* stimuli persisted through the experiment even as listeners were changing their weight of F0 in categorizing Talker 1's *beer/pier* stimuli. The ANOVA for Talker 2's *deer/tear* test showed significant main effects of Block and F0, but no significant Block*F0 interaction [Block: $F(3, 33) = 8.730$, $p < .001$; F0: $F(1, 11) = 10.242$, $p = .008$]. Thus, unlike *beer/pier* stimuli, listeners used F0 information in categorizing new talker's *deer/tear* stimuli. However, listeners did not modulate the use of F0 in the categorization of the new talker's stimuli as they did for categorization of the familiar talker's stimuli. This indicates that learning of reverse F0/VOT correlation with one talker does not necessarily generalize to a new talker. Listeners maintained a separate weighting of F0 in voicing perception for each of the two talkers.

3. Experiment 2

In this experiment, we test whether learning occurs at a general level of "stop voicing" or at a specific level of the phonetic category. Specifically, we investigate whether listeners track competing input correlations simultaneously across two pairs (i.e., *beer-pier* and *deer-tear*), or whether they aggregate information across voicing pairs.

3.1. Methods

Thirty three normal-hearing, native-English listeners were randomly assigned to Group 1 (N=15) or Group 2 (N=18).

The Talker 1 stimuli of Experiment 1 were used. As in Experiment 1, a categorization task examined baseline influence of F0. In the subsequent word recognition task, the F0/VOT correlation in the stimuli shifted in opposing directions for *beer/pier* and *deer/tear* across three exposure/test blocks. Listeners in Group 1 heard *beer/pier* with the canonical English F0/VOT correlation and *deer/tear* with the reversed F0/VOT correlation in Block 1. The correlation shifted to reversed for *beer/pier* and to canonical *deer/tear* in Block 2, and then back to canonical for *beer/pier* and reversed for *deer/tear* in Block 3. This pattern was reversed for Group 2.

If listeners track the F0/VOT correlations separately for bilabials (/b, p/) and alveolars (/d, t/), F0 down-weighting will appear only for the pair for which the input correlation is reversed. If listeners track F0/VOT correlation for a general "voicing" category, thus aggregating distributional statistics across opposing correlations across *beer/pier* and *deer/tear*, there is no correlation overall between F0 and VOT. In this case, we should expect no modulation by F0 across blocks in either group.

The 20 exposure stimuli (open symbols, Figure 1) were presented 10 times per block in random order. The 4 VOT-neutral test stimuli (filled symbols) were each presented 10 times per block, interspersed randomly among the exposure stimuli. The apparatus and procedure were identical to Experiment 1.

3.2. Results

A 9 x 2 (VOT x F0) ANOVA on percent voiceless responses from baseline categorization test, run separately for *beer/pier* and *deer/tear* for each of Group 1 and 2 indicated significant main effects of VOT and F0, and a significant VOT x F0 interaction ($p < .001$ for all). Post-hoc tests indicated that F0 effect was significant at VOT steps 4, 5 and 6, or VOT steps 4 and 5 ($p < .006$ for all). This shows that both groups showed F0 effect in *beer/pier* and *deer/tear* categorization in general, and in particular for the critical stimuli (step 5) used in the subsequent word recognition test.

Figure 3 reports the mean percent voiceless responses for test stimuli across baseline and three blocks for Group 1 (left) and Group 2 (right). In general, the results support the conclusion that dimension-based statistical learning is category-specific, rather than operating at the level of "voicing.

A 4 x 2 (Block x F0) ANOVA was run on percent voiceless response for each of *beer-pier* and *deer-tear* tests for Group 1 and Group 2. The ANOVA for Group 1's *beer/pier* indicated significant main effects of Block and F0, and a significant Block x F0 interaction [Block: $F(3, 42) = 2.872$, $p = .048$; F0: $F(1, 14) = 31.850$, $p < .001$; Block*F0: $F(3, 42) = 2.883$, $p = .047$]. Paired sample t-tests indicated that the F0 effect in the *beer/pier* test was significant in baseline, and in Blocks 1 and 3, when the input correlation was canonical ($p < .013$, alpha adjusted for 4 comparisons). The ANOVA for Group 1's *deer-tear* indicated a significant main effect of F0 and a significant Block x F0 interaction [F0: $F(1, 14) = 32.367$, $p < .001$; Block*F0: $F(3, 42) = 4.694$, $p < .001$]. Paired sample t-tests indicated that the F0 effect in the *deer/tear* test was significant in baseline and in Block 3 (canonical correlation)

($p < .013$, alpha adjusted for 4 comparisons). The F0 effect modulated separately for *beer/pier* and *deer/tear* corresponding to the different patterns of F0/VOT correlation in the input of *beer/pier* and *deer/tear* exposure stimuli.

The ANOVA for Group 2's *beer-pier* indicated a significant main effect of Block and a significant Block*F0 interaction [Block: $F(3, 51) = 2.308$, $p = .009$; F0*Block: $F(3, 51) = 3.113$, $p = .03$]. Paired sample t-tests indicated that F0 effect in *beer/pier* test was significant only in baseline ($p < .013$, alpha adjusted for 4 comparisons). The ANOVA for Group 2's *deer/tear* indicated a significant main effect of F0 and a significant interaction between Block and F0 [F0: $F(1, 17) = 25.173$, $p < .001$; Block*F0: $F(3, 51) = 3.971$, $p = .013$]. Paired sample t-tests indicated that F0 effect in *deer/tear* test was significant in baseline and in Block 1 and 3, when the input correlation was canonical ($p < .013$, alpha adjusted for 4 comparisons).

These results demonstrate that listeners can track separate statistics across speech categories. Group 1 listeners ceased to rely on F0 in recognizing *beer* and *pier* in Block 3 in response to the reverse F0/VOT correlation in the *beer/pier* input, while they maintained reliance on F0 in recognizing *deer* and *tear* in the same block. The same listeners ceased to rely on F0 in recognizing *deer* and *tear* in Block 2 and 4 in response to the reverse F0/VOT correlation in the *deer/tear* input, while maintaining F0 reliance in recognizing *beer* and *pier* in the same blocks.

Group 2 listeners, on the other hand, ceased to rely on F0 in recognizing *beer* and *pier* in all of the three exposure blocks, while they showed modulation of F0 in recognizing *deer* and *tear* in response to the shift of input F0/VOT correlation. These listeners demonstrate separate processing of /b,p/ and /d,t/ categories in Blocks 2 and 4, in which they ceased to rely on F0 in recognizing *beer* and *pier* while maintaining the use of F0 in recognizing *deer* and *tear.*

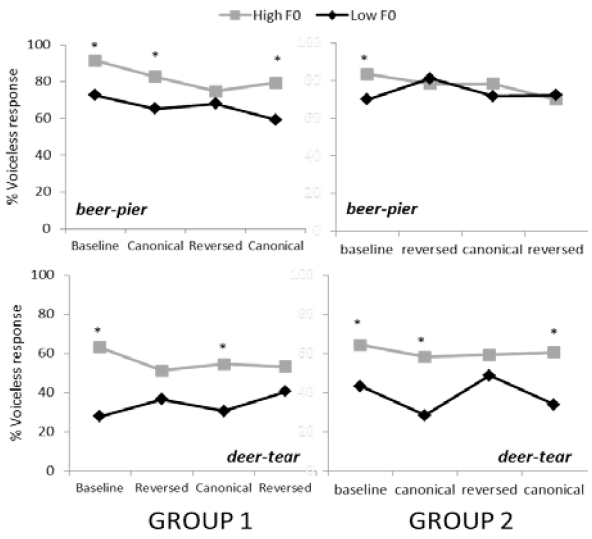

Figure 3: *Mean percent voiceless response across four blocks for Group 1 and Group 2. A star indicates statistical significance.*

4. General discussions

Listeners are sensitive to local acoustic statistics and use them to dynamically "tune" long-term representations by tracking dimensional relationships in online speech processing [3]. Reliance on the dimensions defining perceptual categories (e.g., F0, VOT) is dynamically and rapidly adjusted in online speech processing to accommodate regularities experienced in the immediate speech environment. The current findings suggest that dimension-based statistical learning is specific to talker and phonetic categories.

A previous study [4] demonstrated that dimension-based statistical learning generalized to a new (resynthesized) voice. However, in that case, the new test voice was created from the voice of the speaker listeners experienced during exposure. Although the formant frequencies were shifted to create the impression of new voice by simulating a change in vocal tract size, there may have been subtlely similar source qualities (e.g., relative amplitude of formants and intervals between formants). In the current study, in which the new test sounds were created from an entirely different speaker (Exp 1), we did not observe generalization of learning. The current study provides evidence that when voices are from different speakers, listeners do not readily transfer newly acquired short-term learning with one voice to another. On the one hand this finding suggests that learning is strictly talker specific. But in relation to previous research [4], it leaves open the possibility that it is dependent on similarities across voices.

However, it is clear that learning is specific to individual speech categories. Listeners are sensitive to separate input statistics differentially defining voicing for /b-p/ (bilabial) and /d-t/ (alveolar) (Exp 2). Even though the pairs share the feature of manner of articulation (i.e., stop) and are typically contrasted together as voiceless (p, t) versus voiced stops (b, d), dimension-based statistical learning does not seem to operate at the level of "voicing." Instead it is specific to the details of experienced regularities of phonetic categories. It is also noted that this differential tracking of input statistics occurred even though the sounds were produced by the same speaker.

5. Conclusion

Listeners rely on local input regularities to dynamically "tune" long-term representations by tracking dimensional relationships in online speech processing. Relatively more reliable perceptual sources of information (unambiguous VOT) may adjust perception of less-reliable sources (F0). The current findings suggest that dimension-based statistical learning may be an experience-contingent process, specific to the talker and experienced regularities of phonetic categories.

6. References

[1] Abramson, A. S., & Lisker, L. 1985. Relative power of cues: F0 shift versus voice timing. *Phonetic linguistics: Essays in honor of Peter Ladefoged*, 25–33.

[2] 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.

[3] Idemaru, K., & Holt, L. L. 2011. Word recognition reflects dimension-based statistical learning. *Journal of Experimental Psychology: Human Perception and Performance,* 37(6), 1939-1956.

[4] Holt, L. L., & Idemaru, K. 2011. Generalization of dimension-based statistical learning. *Proceedings of the 17th International Congress of Phonetic Sciences, Hong Kong, China, 882-885.*

[5] Boersma, P. & Weenink, D. 2010. Praat: doing phonetics by computer [Computer program]. Version 5.0, retrieved from *http://www.praat.org/*.