Generalization of dimension-based statistical learning

Lori L. Holt and Kaori Idemaru
Carnegie Mellon University, University of Oregon
loriholt@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 learning in online word recognition, testing generalization of learning across lexical contexts, and testing the extent to which variability in the training inventory affects learning. The results indicate that learning generalizes poorly across lexical contexts, but generalization may be boosted when listeners experience the dimensional regularity across multiple lexical items.

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

1. INTRODUCTION

Speech categories are characterized by multiple acoustic dimensions, some of which carry more information in signalling category affiliation than others, e.g., [1]. Adult listeners exhibit reliable perceptual weights that reflect regularities of acoustic dimensions in the native language, e.g., [4], that are acquired across a long developmental course, e.g. [6]. Perceptual cue weight is the key characteristic of mature phonetic categorization.

Although perceptual cue weighting reflects long-term experience with acoustic regularities of native-language speech input, adult listeners also rapidly adjust perceptual weights in response to deviations from long-term regularities, such as in experiencing a foreign accent. Previous research [3] showed that when the long-term English relationship of fundamental frequency (F0) to voicing categories, as implemented in *beer* versus *pier* and *deer* versus *tear*, reversed in the local, short-term speech input, listeners adjusted the perceptual weight of F0 in judging the voicing categories. Subsequent research [2] further demonstrated that this rapid adjustment of perceptual cue weighting is very specific to the phonetic category for which short-term regularities are experienced. When the long-term English relationship of F0 to VOT was reversed in *beer* and *pier* but maintained in *deer* and *tear* in short-term experience, listeners' down-weighting of reliance on F0 occurred only in categorizing *beer* and *pier* while simultaneously using F0 in categorizing *deer* and *tear* (and vice versa).

These studies demonstrate the flexibility of perceptual cue weighting as a function of local input statistics at the level of fine-grained phonetic dimensions and its specificity as it resists generalization across categories that are commonly grouped together, such as stops for example. However, the previous research tested learning and its generalization with just four stimulus words; this is a very conservative experimental test of the information needed to evoke learning. Investigating the various conditions under which generalization does or does not occur is important in understanding the nature of the learning.

The current study presents two experiments. Experiment 1 tested the specificity of learning with exposure to a short-term deviation in *beer*/*pier* tested also across *bear*/*pear*. Experiment 2 tested whether experiencing the short-term deviation across a larger inventory, including lexical and nonce words *beer*/*pier, bill*/*pill, best*/*pest,* and *borth/porth,* affected generalization of learning to *bear/pear.*

2. EXPERIMENT 1

2.1. Methods

Thirty-two native-English listeners with normal hearing participated.

2.1.1. Stimulus creation

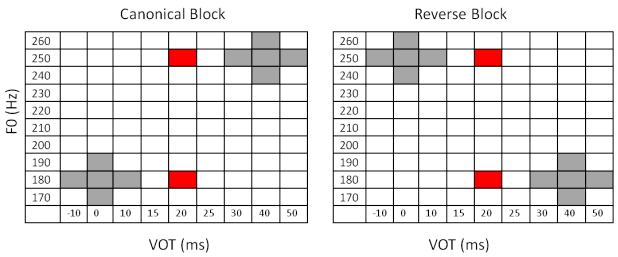
Natural utterances of *beer, pier, bear* and *pear* were digitally recorded (22.05 kHz) in a sound attenuated booth by adult female native speaker. The pairs of end-points were selected for similar duration (385 ms) and F0 contour, and normalized to the same root-mean-square amplitude. Using progressive cross-splicing [5] of these tokens, *beer-pier* and *bear–pear* continua were created with VOT values of 0, 10, 15, 20, 25, 30, 40, and 50 ms. Sounds with -10ms of VOT were created by taking 10 ms of pre-voicing in the voiced production of the same speaker and inserting it before the burst of the voiced endpoint token (VOT = 0ms).

The F0 of the two series (*beer-pier* and *bear-pear*) were then manipulated such that the onset frequency of the vowel was adjusted from 170 Hz to

190 Hz (low F0s), and from 240 Hz to 260 Hz (high F0s) in 10-Hz steps. For each stimulus, the F0 contour of the original production was manually manipulated using Praat 5.3 to adjust the target onset F0 values. From the onset, the F0 decreased quadratically to 150 Hz. The high and low values of F0 and the contour modelled the natural production of the speaker.

From the resulting continua, stimuli illustrated in Fig 1 with colored cells were chosen for the experiment. Fig 1 illustrates characteristics of stimuli with regards to the VOT of the stop and onset F0 of the following vowel. Those indicated by gray cells were exposure stimuli, and red squares were test stimuli. As the experiment tests the generalization of learning from *beer/pier* to *bear/pear*, exposure (gray) stimuli included only *beer/pier* series, whereas the test stimuli (red) consisted of both *beer/pier* (exposed test stimulus pair) and *bear/pear* (generalization test pair).

Figure 1: Schematic illustration of stimulus sampling in the Canonical blocks (left) and the Reverse block (right) as a function of VOT and F0.



2.1.2 Procedure

In the first block (Canonical1), listeners heard exposure stimuli, *beer* and *pier,* with the Canonical English F0/VOT relationship: voiced stops had lower F0s whereas voiceless stops had higher F0s in the following vowel (Fig 1). In the second block (Reverse), listeners heard *beer* and *pier* with the F0/VOT correlation reversed: voiced stops were associated with higher F0s and voiceless stops with lower F0s. In the last block (Canonical 2), the F0/VOT correlation returned to canonical. In each block, the exposure stimuli (gray cells) were presented 20 times each in random order, for a total of 200 exposure trials. All exposure stimuli were *beer* and *pier*. The VOT-neutral test stimuli (red cells) were VOT neutral *beer/pier* and *bear/pear* tokens, each of which was presented 10 times per block. They were interspersed randomly among the exposure stimuli, for a total of 40 test trials per block. Trials proceeded continuously across the three blocks as listeners performed a two-alternative word-recognition task in a sound attenuated booth. 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. In this and the subsequent experiments, there were a total of 600 exposure trials and 120 test trials. The entire session was completed in approximately 20 minutes.

2.2 Results

Fig 2 shows proportion voiceless response for the High F0 and Low F0 VOT-neutral test stimuli for Experienced and Generalization conditions. Prior to the exposure test, listeners categorized *beer-pier* and *bear-pear* continua varying along 9-step VOTs (Fig 1) and low (180Hz) and high (250 Hz) F0s to verify the baseline F0 effect in voicing categorization. The responses to the VOT neutral stimuli (VOT = 20ms) from the categorization test are plotted as Baseline in Fig 2.

A 2 (generalization: exposed vs. generalization) x 4 (blocks) x 2 (F0: high vs. low) repeated-measure ANOVA on the mean proportion voiceless responses showed a significant main effect for Generalization and F0, as well as significant Generalization x F0, Block x F0, and Generalization x Block x F0 interactions ($p < .001$ for each). Given the significant Generalization x Block x F0 interaction, 4 (Block) x 2 (F0) ANOVAs were run separately for the Experienced and the Generalization condition to examine the across-block modulation of F0 effect in each condition.

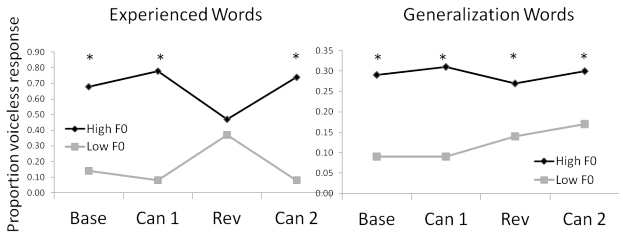
In the Experienced condition, the Block x F0 interaction, $F(3, 93) = 57.222$, $p < .0001$, $\eta^2 = .649$, as well as the main effect of F0, $F(1, 31) = 280.434$, $\eta^2 = .900$, were statistically significant, indicating the F0 effect modulated across blocks. Comparisons of the proportion voiceless response between High and Low F0 conditions in each block indicated that the difference was significant in all blocks ($p < .013$ for all, with alpha adjusted for 4 comparisons) except for the Reverse block. These results replicated the prior findings [2, 3] that when listeners experience a short-term reversal of F0/VOT correlation in /b/ and /p/ in *beer* and *pier*, they down-weight reliance on F0 in categorizing VOT neutral /b-p/ in the experienced *beer*/*pier* context.

In the Generalization condition, the ANOVA showed a significant main effect of F0, $F(1, 31) = 47.322$, $p < .001$, $\eta^2 = .604$, but there was no significant Block x F0 interaction, indicating that the effect of F0 persisted across the 4 experiment blocks for the Generalization condition (*bear/pear*) without modulation of the effect. A post-hoc analysis was conducted with a 2 (Block: Canonical 1 vs. Reverse) x 2 (F0) ANOVA to explore a localized change in F0 effect when the F0/VOT correlation reversed.

The results showed a Block x F0 interaction that approached statistical significance, $F(1, 31)=3.866$, $p = .058$, $\eta^2 = .088$. This suggests that generalization may have occurred but the current tests are not sensitive enough to detect it.

These results indicate that the robust short-term learning, down-weighting of reliance on F0, that occurred in /b-p/ categorization in *beer* and *pier* did not generalize with the same or similar magnitude to /b-p/ categorization in new lexical contexts, *bear* and *pear,* with which listeners did not experience the short-term F0/VOT reversal. The data does suggest possible generalization, but if it occurred, the effect of learning is considerably smaller. The scope across which learning operates (for example whether it is a lexical or syllabic or acoustic) is yet to be determined.

Figure 2: Proportion voiceless response for high F0 and low F0 test stimuli across 4 experiment blocks for Experienced (left) and Generalization (right) words. Blocks are Baseline, Canonical 1, Reverse, and Canonical 2.



3. EXPERIMENT 2

3.1. Methods

Thirty-two native-English listeners with normal hearing participated. Listeners experienced the short-term F0-VOT correlation reversal in *multiple* words: *beer, pier*, *bill, pill, best, pest,* non-word *borth* and *porth*, and were tested with *beer/pier* and *bear/pear* test stimuli to determine whether experiencing the correlation reversal across a larger word list influences generalization.

The stimulus construction method and procedure were the same as Exp 1. In Exp 2, exposure (gray) stimuli included *beer/pier, bill/ pill, best/pest,* and *borth*/*porth* series, and the test stimuli (red) consisted of *beer/pier* (Experienced) and *bear/pear* (Generalization), possessing the values of VOT and F0 as illustrated in Fig 1. The experiment proceeded from Canonical 1, to Reverse and to Canonical 2 block, with the block structure implicit to the participants, and each block containing 200 exposure and 40 test trials (same as Exp 1).

3.2. Results

A 2 (Generalization) x 4 (Blocks) x 2 (F0) repeated-measure ANOVA returned a significant main effect for all factors, and all interactions were also significant ($p < .001$ for each). Given a significant Generalization x Block x F0 interaction, $F(3, 93) = 3.005$, $p < .05$, $\eta^2 = .088$, separate 4 (Block) x 2 (F0) ANOVAs were run for the Experienced and the Generalization conditions. Both tests showed a significant main effect for Block and F0 ($p < .001$ for both), and a significant interaction between the two: $F(3, 93) = 14.177$, $p < .001$, $\eta^2 = .314$ in the Experienced condition, and $F(3, 93) = 3.392$, $p < .05$, $\eta^2 = .099$ in the Generalization condition, indicating that the influence of F0 on voicing categorization modulated across blocks whether the words were experienced words or generalization words. The significant Generalization x Block x F0 interaction in the initial ANOVA, as well as the significant Block x F0 interaction for each of Experienced and Generalization conditions, suggest that the interaction (modulation of F0 influence across blocks) was present in both conditions, but the magnitude of the interaction varied across them. The difference in the effect size of the interaction for the Experienced condition ($\eta^2 = .314$) and the Generalization condition ($\eta^2 = .099$) was noted.

The presence of dimension-based statistical learning while recognizing Experienced words was confirmed by post-hoc tests. The first set of tests comparing proportion voiceless response between High and Low F0 conditions indicated a significant difference in all blocks (Fig 3, left panel), unable to locate the modulation of F0 influence. Another set of tests comparing the extent of F0 influence, computed as the difference in proportion voiceless response (difference score) due to High and Low F0, across blocks (Fig 3, right panel) showed F0 influence decreased from Baseline and Canonical 1 to Reversed ($p <.008$ for both, alpha adjusted for 6 comparisons), although its increase from Reverse to Canonical 2 was not significant ($p = .056$, alpha adjusted to .008 for 6 comparisons). Although the use of F0 did not bounce back in the final Canonical 2 block, down-weighting of F0 in the Reverse block confirms the dimension-based statistical learning.

Generalization of this learning to a new lexical context was also evident. Comparisons of proportion voiceless responses between High and Low F0 conditions in each block (Fig 4, left panel) indicated that the difference was significant in Baseline and Canonical 1 ($p < .013$ for both, with alpha adjusted for 4 comparisons) but not in the Reverse block. The comparison in Canonical 2 approached statistical significance ($p = .018$, alpha adjusted to .013 for 4

comparisons). These indicate that listeners' reliance on F0 in voicing categorization disappeared in the reverse block in recognizing generalization words (*bear/pear*), with which listeners did not experience the short-term F0VOT correlation. When the correlation in the input returned to the long-term pattern, the listeners' response showed a trend of returned reliance on F0 for categorization.

Another set of tests comparing the difference score (indicating the extent of F0 influence) across blocks also showed a reliable change in this measure from Baseline to Reverse ($p < .008$, alpha adjusted for 6 comparisons), indicating down-weighting of F0 reliance in the Reverse block. However, no significant change from Canonical 1 to Reversed, or from Reversed to Canonical 2 was obtained.

In this experiment, listeners experienced the reversal of F0/VOT correlation in /b/ and /p/ in multiple lexical contexts: *beer, pier, bill, pill, best, pest,* and non-word *borth* and *porth.* They down-weighted reliance on F0 in voicing categorization not only in recognizing *beer* and *pier,* in which they experienced the F0/VOT reversal, but also in recognizing *bear* and *pear* in which they did not experience the reversal.

5. CONCLUSION

Listeners track acoustic dimensional relationships in online speech processing. In response to a short-term deviation from long-term representations, listeners dynamically "tune" reliance on acoustic dimensions for word recognition. The current findings demonstrate that generalization of this dimension-based statistical learning of /b-p/ categorization to a new lexical context seems very weak (Exp 1). It is, however, yet to be determined whether the generalization was constrained because the target sounds /b-p/ occurred in a different lexical context (e.g., *bear* as opposed to the experienced *beer*), in a different vowel context (e.g., /bɛ/ as opposed to the experienced /bɪ/), or with different acoustics (e.g., an instance of /b/ in *bear* as opposed to the experienced and acoustically different /b/ in *beer*). The current findings also demonstrate that experiencing a short-term deviation in multiple contexts may foster generalization (Exp2). However, the effect of generalization is also weak in this case. Future research is needed to examine at which level of context (e.g., syllabic, lexical, or acoustic) and the nature of the context that may foster generalization.

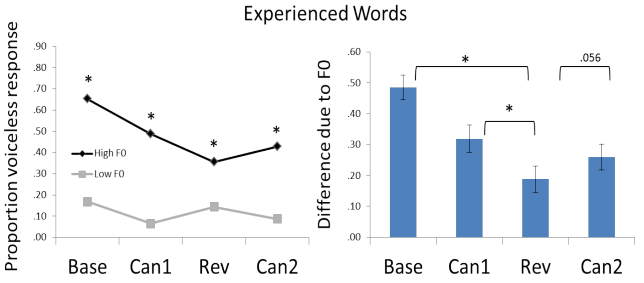


Figure 3: Proportion voiceless response for high F0 and low F0 test stimuli (left) and difference score due to F0 (right) for Experienced words.

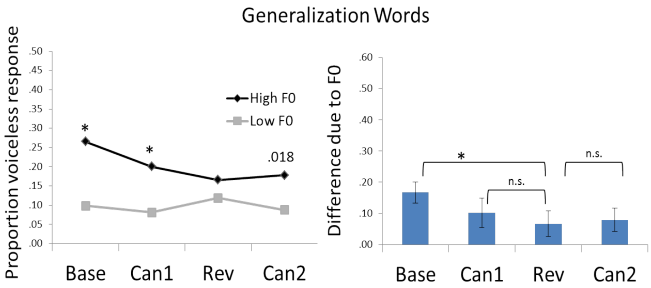


Figure 4: Proportion voiceless response for high F0 and low F0 test stimuli (left) and difference score due to F0 (right) for Generalization words.

5. 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] Idemaru, K., Holt, L. L. 2014. Specificity of dimension-based statistical learning. *J. Exp. Psychol.-Hum. Percept. Perform.*, *40*(3), 1009-1021.

[3] Idemaru, K., Holt, L. L. 2011. Word recognition reflects dimension-based statistical learning. *J. Exp. Psychol.-Hum. Percept. Perform.*, 37(6), 1939-1956

[4] Lotto, A.J., Sato, M., Diehl, R.L. 2004. Mapping the task for the second language learner: The case of Japanese acquisition of /r/ and /l/. In: Slifka, J., Manuel, S., Matthies, M. (eds), *From Sound to Sense: 50+ Years of Discoveries in Speech Communication*, 181-186.

[5] McMurray, B., Aslin, R. N. (2005). Infants are sensitive to within-category variation in speech perception. *Cognition, 95*(2), B15-B26.

[6] Nittrouer, S. 1992. Age-related differences in perceptual effect of formant transitions within syllables and across syllable boundaries. *J. Phon*, *20,* 1–32.