Journal of Experimental Psychology:
Human Perception and Performance
2014, Vol. 40, No. 3, 1009–1021

0096-1523/14/$12.00 DOI: 10.1037/a0035269

Specificity of Dimension-Based Statistical Learning in Word Recognition

Kaori Idemaru
University of Oregon

Lori L. Holt
Carnegie Mellon University

Speech perception flexibly adapts to short-term regularities of 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 a fixed value along the dimension. Two experiments examine the nature of this *dimension-based statistical learning* in online word recognition, testing generalization of learning across phonetic categories. While engaged in a word recognition task guided by perceptually unambiguous voice-onset time (VOT) acoustics signaling stop voicing in either bilabial rhymes, *beer* and *pier*, or alveolar rhymes, *deer* and *tear*, listeners were exposed incidentally to an artificial "accent" deviating from English norms in its correlation of the pitch onset of the following vowel (F0) with VOT (Experiment 1). Exposure to the change in the correlation of F0 with VOT led listeners to down-weight reliance on F0 in voicing categorization, indicating dimension-based statistical learning. This learning was observed only for the "accented" contrast varying in its F0/VOT relationship during exposure; learning did not generalize to the other place of articulation. Another group of listeners experienced competing F0/VOT correlations across place of articulation such that the global correlation for voicing was stable, but locally correlations across voicing pairs were opposing (e.g., "accented" *beer* and *pier*, "canonical" *deer* and *tear*, Experiment 2). Listeners showed dimension-based learning only for the accented pair, not the canonical pair, indicating that they are able to track separate acoustic statistics across place of articulation, that is, for /b-p/ and /d-t/. This suggests that dimension-based learning does not operate obligatorily at the phonological level of stop voicing.

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

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 (e.g., Clayards, Tanenhaus, Aslin, & Jacobs, 2008; Norris, McQueen, & Cutler, 2003; Eisner & McQueen, 2005; Kraljic & Samuel, 2006, 2007; Maye, Werker, & Gerken, 2002; Reinisch & Holt, 2013; Idemaru & Holt, 2011).

For example, speech processing rapidly adjusts the perceptual weight of acoustic dimensions defining speech categories in response to perturbations of long-term regularities experienced in short-term input. Idemaru and Holt (2011) presented listeners with artificially "accented" rhymes *beer*, *pier*, *deer* or *tear*, in which the correlation between two critical acoustic cues to voicing categorization, voice onset time (VOT) and fundamental frequency (F0) of the vowel following the stop, was manipulated. When the correlation between VOT and F0 in *beer*, *pier*, *deer* and *tear* was reversed from the English norm (Abramson & Lisker, 1985; Idemaru & Holt, 2011; higher F0s were paired with voiced stops [*beer* and *deer*] and lower F0s were paired with voiceless stops [*pier* and *tear*]), listeners down-weighted reliance on F0 within just a few trials of exposure such that it no longer influenced voicing categorization.

These results demonstrate that listeners track relationships between acoustic dimensions in online speech processing and that the diagnosticity of an acoustic dimension to phonetic category membership is not simply a fixed function of its value along the acoustic dimension. Rather, it is evaluated relative to evolving local regularities between acoustic dimensions experienced across short-term experience. This perceptual tuning is likely to be important for understanding how listeners deal with the acoustic perturbations to speech resulting from adverse listening conditions arising from accent, dialect and dysarthria. Idemaru and Holt (2011) referred to this as *dimension-based statistical learning* to highlight that, in addition to being sensitive to regularities among perceptual or linguistic "objects" like syllables or words (e.g., Saffran, Aslin, & Newport, 1996; Newport & Aslin, 2004), listen-

This article was published Online First December 23, 2013.

Kaori Idemaru, Department of East Asian Languages and Literatures, University of Oregon; Lori L. Holt, Department of Psychology and Center for the Neural Basis of Cognition, Carnegie Mellon University.

We thank Christi Gomez, Rio Omachi, and Heather Cates for running the experiments. This research was supported by the National Institutes of Health (R01DC004674), the National Science Foundation (0746067), and Research, Innovation, and Graduate Education, University of Oregon. A portion of this work was presented at Phonological Forum, Japan, August, 2012, the 14th Australasian International Conference on Speech Science and Technology, Australia, December 2012, and at the 17th International Congress of Phonetic Sciences, China, August, 2011.

Correspondence concerning this article should be addressed to Kaori Idemaru, Department of East Asian Languages and Literatures, University of Oregon, Eugene, OR 97403. E-mail: idemaru@uoregon.edu

ers also track regularities among physical dimensions that define such objects and use this information to constrain online perception.

These findings are situated in a growing literature indicating that multiple information sources including lexical (Norris, McQueen, & Cutler, 2003; Eisner & McQueen, 2005; Kraljic & Samuel, 2006, 2007; Maye, Werker, & Gerken, 2002; Reinisch & Holt, 2013), visual (Bertelson, Vroomen, & de Gelder, 2003; Vroomen, van Linden, de Gelder, & Bertelson, 2007), phonotactic (Cutler, McQueen, Butterfield, & Norris, 2008), and statistical (Idemaru & Holt, 2011; Clayards et al., 2008) information support fairly rapid, online adjustments to phonetic categorization in response to deviations of speech from the norm. Although each of these sources of information may drive phonetic retuning effects, it is not yet clear whether they rely upon common mechanisms. In lexically guided phonetic retuning, for example, top-down feedback from lexical knowledge serves to tune how the system encodes incoming speech when ambiguous speech sounds are embedded in lexical contexts for which only one phonetic alterative forms a real word (e.g., an ambiguous /d/-/t/ sound is heard as /d/ in the context of *avoca__o*, but as /t/ in *luna__ic*; see McClelland, Mirman, & Holt, 2006; Mirman, McClelland, & Holt, 2006; Norris et al., 2003 for debate on the details of how this may occur). Dimension-based statistical learning (Idemaru & Holt, 2011) differs from lexically guided phonetic tuning in that lexical information does not disambiguate the ambiguous speech acoustics; all phonetic possibilities are real words (*deer/tear, beer/pier*). Thus lexical status could not serve as a teaching signal to drive learning.

Rather than a lexical "teacher," Idemaru and Holt (2011) argue that the reliable, unambiguous VOT information experienced throughout exposure to the shifting correlation between F0 and VOT dimensions served as a signal to orient the relationship of the secondary, F0, dimension to voicing categories. In light of the many information sources that are effective in tuning phonetic categorization, Idemaru and Holt (2011) question whether any consistent source of information (i.e., not only higher-order feedback) may be exploited as a "teacher" signal to drive adaptive plasticity in speech processing.

What remains unclear is the extent to which the learning resulting in phonetic retuning that arises from experiencing disambiguating information across various information sources share commonalities. Understanding the nature of this adaptive plasticity in speech perception is enhanced by investigations of how learning generalizes. This approach has been especially productive with respect to lexically guided phonetic retuning (Eisner & McQueen, 2005; Kraljic & Samuel, 2006; Maye et al., 2002; Reinisch & Holt, 2013). In the present study, we take this approach to examine generalization of dimension-based statistical learning.

Using generalization patterns, we examine the extent to which dimension-based statistical learning for English listeners' perceptual weighting of F0 in the context of VOT operates at the level of individual stop consonant categories specific to place-of-articulation (e.g., voicing at bilabial vs. alveolar as in *beer/pier* and *deer/tear*) versus at the level of phonological voicing class across stop place of articulation (voicing in general as in *beer/deer* vs. *pier/tear*). If the latter is true then we predict learning at one place of articulation (e.g., *beer/pier*) will generalize to the other place (*deer/tear*). If, however, listeners track the relationship between F0 and VOT dimensions independently across place of articulation, generalization may not be evident. The current work thus investigated generalization of dimension-based statistical learning across stop places of articulation (Experiment 1), with results suggesting learning specific to place of articulation. The next experiment examined the extent of this learning's specificity, investigating whether listeners can simultaneously learn opposing statistical patterns across stop places of articulation, or whether global statistical patterns at the level of voicing dominate when place-level statistics compete (Experiment 2).

Experiment 1

In Experiment 1, one group of listeners (BP exposure group) experienced an artificial accent reversing the F0/VOT correlation imposed on stop voicing only among the bilabial stops (i.e., *beer-pier*), and another group (DT exposure group) experienced the accented production only among the alveolar stops (i.e., *deer-tear*). Both groups were tested for adjustment to the accent with *beer-pier* and *deer-tear* test stimuli to examine generalization of dimension based statistical learning across place of articulation.

Method

Participants. Twenty-seven native-English listeners with normal hearing participated. They were either university students or employees. Participants were randomly assigned to a BP exposure group ($n = 14$) or a DT exposure group ($n = 13$).

Stimulus creation. Stimuli from Idemaru and Holt (2011) served as the stimuli in this experiment. The stimuli were created based on natural utterances of *pier* and *tear* produced in isolation by a female monolingual native speaker of midwest American English (second author). Using these utterances as end points, 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. These ranges were chosen based on a pilot categorization test indicating category boundaries at about 10-ms VOT for the *beer-pier* series and 20-ms VOT for the *deer-tear* series for this speaker. The shift in voicing category boundary with place of articulation is typical of English voicing perception (Abramson & Lisker, 1985).

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 (Boersma & Weenink, 2010). The first 10 ms of the original voiceless productions were left intact to preserve the consonant bursts. For the negative VOT values, prevoicing 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 fundamental frequency (F0) was manipulated such that the F0 onset frequency of the vowel, [**I**], following the stop consonant was adjusted from 220 Hz to 300 Hz across nine 10-Hz steps. These F0 values were determined based on the minimum voiced and maximum voiceless F0 values (approximately 230 Hz for voiced and 290 Hz for voiceless) of the speaker across multiple productions of the stimulus words. For each stimulus, the F0 contour of the original production was 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. This contour was modeled on this speaker's natural productions.

Baseline categorization stimuli. Before the exposure and test to index dimension-based statistical learning, listeners categorized rhymes *beer-pier* and *deer-tear* stimulus continua varying in the VOT and F0 dimensions to measure the baseline influence of F0 on voicing judgments. These stimuli varied along VOT in seven 10-ms steps (from −20 ms to 40 ms for *beer-pier* and from −10 ms to 50 ms for *deer-tear*) and along F0 in two levels (230 Hz and 290 Hz). Stimuli were presented 10 times each, blocked for *beer-pier* and *deer-tear* with the block order counterbalanced across participants.

Exposure stimuli. Exposure stimuli had perceptually *unambiguous* VOT values differentiating the voicing categories; however, the relationship between the VOT and F0 changed across the course of the experiment, exposing listeners to a short-term deviation in the F0/VOT correlation typical of English voicing categories. Figure 1 illustrates the two-dimensional F0/VOT acoustic space from which stimuli were drawn. The stimuli indicated by large symbols were used in the experiment; open symbols indicate exposure stimuli, whereas filled symbols were test stimuli (see below).

In previous experiments (Idemaru & Holt, 2011), participants were exposed to the shift of F0/VOT correlation from the canonical English pattern (higher F0 for voiceless stops) (e.g., Abramson & Lisker, 1985) to the reversed pattern (lower F0 for voiceless stops) for both *beer-pier* and *deer-tear* stimuli. In the current experiment listeners experienced the F0/VOT correlation shift at only one stop place of articulation: the BP exposure group heard only *beer* and *pier* (bilabial) exposure stimuli with the reversed F0/VOT correlation, and the DT exposure group heard only *deer* and *tear* (alveolar) exposure stimuli with the reversed correlation. Listeners did not experience the artificial "accent" with the reversed F0/VOT correlation for the unexposed place of articulation; for the unexposed place of articulation, they heard only test stimuli and did not experience exposure stimuli for which F0 covaried with perceptually unambiguous VOTs.

In the first block, listeners heard the designated voicing pair with the familiar canonical English F0/VOT correlation: voiced stops had lower F0s whereas voiceless stops had higher F0s (e.g., Abramson & Lisker, 1985). For example, the BP exposure group heard *beer* with lower F0s and *pier* with higher F0s. In these canonical correlation exposure stimuli, perceptually *unambiguous* short VOT values (−20, −10, and 0 ms for the *beer-pier* series, heard as [b]; −10, 0, and 10 ms for the *deer-tear* series, heard as [d]) were combined with low F0s (220, 230, and 240 Hz), whereas long VOT values (e.g., 20, 30, and 40 ms for the *beer-pier* series, heard as [p]; 20, 40 and 50 ms for the *deer-tear* series, heard as [t]) were combined with high F0s (280, 290, and 300 Hz). These VOT and F0 values were not fully crossed for each category (e.g., /b/) so that the category center values of VOT and F0 occurred more frequently than peripheral values in the resulting five stimuli within each category (see Figure 1).

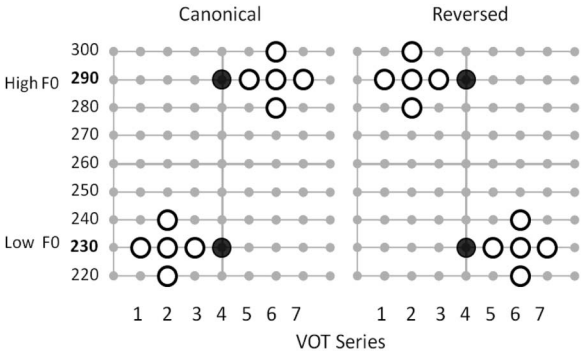

Figure 1. F0/VOT correlation in stimuli across Canonical and Reversed experimental blocks. Large open symbols are exposure stimuli, and large filled symbols are test stimuli.

In the second block, the F0/VOT correlation was reversed such that listeners heard the designated stops with an F0/VOT correlation opposite their long-term experience with English (Reverse correlation). For example, for the BP exposure group, *beer* was now paired with high F0s (280, 290, and 300 Hz), whereas *pier* was now associated with low F0s (220, 230, and 240 Hz). For each group, listeners experienced the canonical and reversed correlations of F0 and VOT only for their designated voicing pair, *beer-pier* (Condition BP) or *deer-tear* (Condition DT). Note that for these stimuli, VOT unambiguously signaled voicing category; although F0 was correlated with VOT, it was never essential for speech categorization, which could be accomplished entirely with the unambiguous VOT.

Thus 10 unique tokens formed a list of exposure stimuli for each block (open symbols in Figure 1) for either *beer-pier* or *deer-tear*, according to condition. These stimuli were presented in 30 random orders per block to expose listeners to the canonical or reversed F0/VOT correlation. The block structure was implicit in the experimental session, serving only to define the type of stimuli presented. It was not apparent in the nature of the task. Participants were not informed that they were in a particular group, that the experiment was divided into blocks, or that the characteristics of the spoken words would change. Trials proceeded continuously across changes in the relationship of F0 to VOT, and listeners performed the same word recognition task throughout the experiment.

Test stimuli. To assess listeners' sensitivity to changes in the F0/VOT correlation, test stimuli with perceptually *ambiguous* VOT values were interspersed among the exposure stimuli throughout the experiment (see large filled symbols in Figure 1). F0 exerts the strongest influence on voicing perception when VOT is ambiguous (Abramson & Lisker, 1985; Idemaru & Holt, 2011) and thus the VOT-neutral test stimuli provided an opportunity to observe subtle changes in listeners' use of the F0 dimension as a function of experienced changes in the correlation between F0 and VOT. With VOT eliminated as a cue to voicing category, the difference in word recognition of high F0 versus low F0 test stimuli provides a measure of reliance upon F0 in voicing judgments. The VOT-neutral test stimuli for both *beer-pier* and *deer-tear* were presented to both the BP and the DT exposure groups.

The test stimuli were constant across blocks and possessed perceptually ambiguous VOT values (10 ms for *beer-pier* type and 20 ms for *deer-tear* type) with low- and high-F0 frequencies (230 and 290 Hz) corresponding to the midpoint F0 frequencies of the exposure stimuli within these ranges. These four test stimuli (*beer-pier, deer-tear* × 2 F0s) were presented 10 times in each block in a random order interspersed among the exposure stimuli. The test

stimuli were not described to participants, and they were not differentiated from exposure stimuli by task or instructions.

Procedure. A categorization task (7 VOTs × 2 F0s × *beer-pier, deer-tear* × 10 times, a total of 280 trials) examined the baseline effect of F0 before exposure to accented stimuli. This provided a test to confirm that participants' voicing judgments reflected experience with the long-term F0/VOT correlation typical of English, as has been observed in many previous studies (Abramson & Lisker, 1985; Haggard, Ambler, & Callow, 1970; Whalen, Abramson, Lisker, & Mody, 1993; Idemaru & Holt, 2011). Participants were seated in front of a computer monitor in a sound booth. Each trial consisted of a spoken word presented 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, 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 the 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 Canonical and Reversed F0/VOT correlations via exposure stimuli and monitored reliance upon F0 to categorize VOT-neutral test stimuli. The BP group experienced the F0/VOT correlation change only in *beer-pier* exposure stimuli and tested with both *beer-pier* and *deer-tear* test stimuli. The DT group experienced the F0/VOT shift only in *deer-tear* exposure stimuli and tested with both pairs of test stimuli. The 10 unique exposure stimuli were presented 30 times per block, for a total of 600 exposure trials, consistent with Idemaru and Holt (2011). The VOT-neutral test stimuli were each presented 10 times per block, for a total of 40 test trials per block (*beer-pier, deer-tear* × 2 F0s × 10 times).

The procedure and apparatus for this task were identical to those for the baseline categorization task. Listeners in BP group, for example, heard *beer* or *pier* most of the time, and *deer* or *tear* occasionally, and selected their response from the two images (i.e., icons of beer and pier or those of deer and tear) on the monitor by pressing a designated key number. Trials proceeded continuously across the two blocks as listeners performed the two-alternative word-recognition task. The block structure was implicit: participants were not informed that the experiment was divided into separate blocks, or that the nature of the acoustic cues would vary.

Results

Baseline test. A 2 × 7 × 2 × 2 (Place × VOT × F0 × Group) ANOVA with repeated measures on the first three factors was run on the mean percent voiceless responses to the baseline categorization test. The factor of Place refers to the two places of articulation for voicing: *beer-pier* and *deer-tear*. The factor of Group refers to the BP versus DT exposure groups. The test revealed significant main effects of Place, VOT, F0, but not of Group [Place: $F(1, 25) = 48.822$, $p < .001$; VOT: $F(6, 150) = 1375.824$, $p < .001$; F0: $F(1, 25) = 44.668$, $p < .001$; Group: $F(1, 25) = .417$, $p = .524$], as well as a number of significant interactions. Critical to this experiment, a significant VOT × F0 interaction [$F(6, 150) = 19.558$, $p < .001$] indicated that the influence of F0 was modulated by VOT, as expected from the fact that F0 exerts the strongest influence on voicing perception when VOT is ambiguous (Abramson & Lisker, 1985; Idemaru & Holt, 2011). A significant Place × VOT × F0 interaction and nonsignificant Place × VOT × F0 × Group interaction [Place*VOT*F0: $F(6, 150) = 6.738$, $p < .001$; Place*VOT*F0*Group: $F(6, 150) = .266$, $p = .952$] indicated that this modulation of the influence of F0 varied across *beer-pier* and *deer-tear* categorization, and this was consistent across the two groups. This confirmed that two groups were compatible with regard to the use of F0 for *beer-pier* and *deer-tear* categorization before the exposure task.

A planned comparison between high and low F0 stimuli collapsed for the groups at the middle step of the VOT series (step 4) was run for each of the places of articulation as this was the critical VOT value used as the ambiguous VOT test stimulus used later in the experiment. The mean F0 effect (i.e., difference in percent voiceless response between high F0 and low F0) at the middle VOT stimulus was 13.3% ($SE = 3.29$) for *beer-pier* and 30.0% ($SE = 4.70$) for *deer-tear*. The results of the comparisons indicated that this F0 effect at the middle VOT step was statistically significant for voicing categorization at both bilabial and alveolar places of articulation ($p < .001$ for both). This means that before the exposure experiment, listeners in both groups used F0 as a cue to distinguish *beer* from *pier* as well as *deer* from *tear* when the VOT was ambiguous, consistent with expectations from prior research (e.g., Abramson & Lisker, 1985; Idemaru & Holt, 2011). The significant influence of the factor Place was attributable to difference in the magnitude of the effect (i.e., 13.3% for *beer-pier;* 30.0% for *deer-tear*). This magnitude difference of the effect of F0 across the two places of articulation (the data at the baseline in Figure 2) is consistent with previous experiments (see Idemaru & Holt, 2011). The baseline can be thought of as a block with no inherent F0/VOT correlation because stimuli were sampled with equal probability across VOT for High and Low F0s. Thus, the observed effect of F0 at baseline reflects listeners' long-term experience.

Exposure test. Figure 2 reports the mean percent voiceless response for test stimuli across baseline and two exposure blocks for the exposed words (top) and new words (bottom). As is apparent in the figure, the influence of F0 on voicing judgments changed across blocks for exposed words, but this learning did not generalize to new words at a different place-of-articulation.

A 2 × 3 × 2 × 2 (Generalization × Block × F0 × Group) ANOVA with repeated-measures on the first 3 factors was run on the mean percent voiceless responses. The factor of Generalization included two levels, Exposed and Generalization. Note that the content of Exposed and Generalization words was different across BP and DT groups (see Table 1). Namely, the Exposed test words were *beer-pier* for the BP group and *deer-tear* for the DT group, whereas the Generalization test words were *deer-tear* for the BP exposure group and *beer-pier* for the DT exposure group. The factor of Group refers to BP and DT groups. The responses to the VOT-neutral stimuli in the baseline categorization task (the middle Step 4 stimuli) were included in the analysis as one of the three blocks (i.e., baseline, canonical correlation, and reversed correlation).

If there is generalization of dimension-based statistical learning from exposed words to generalization words at a different place of articulation, we would expect a Block by F0 interaction regardless of the factor of Generalization (Exposed or Generalization) or

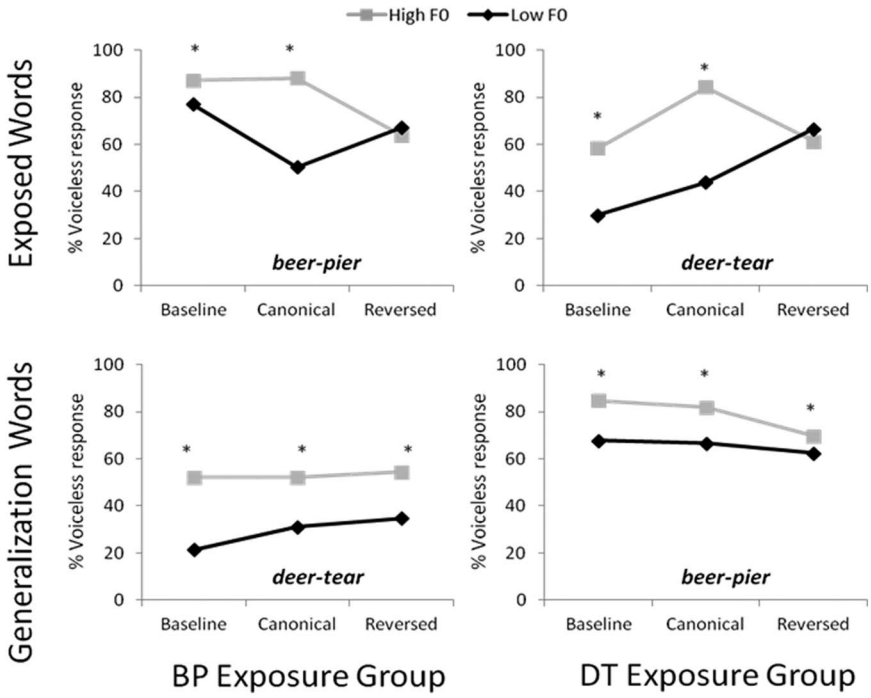

Figure 2. Mean percent voiceless response across three blocks for BP Exposure Group (left) and DT Exposure Group (right) by Exposed Words (top) and New Words (bottom) in Experiment 1. A star indicates statistical significance ($p < .008$).

Group (BP or DT), indicating modulation of the influence of F0 by the F0/VOT correlation (Block) for both exposed and generalization words and by both BP and DT groups. However, if the learning is specific to words for which the F0/VOT correlation shifted, we would expect a Generalization × Block × F0 interaction for both groups.

The test returned significant main effects of Generalization and F0, as well as a number of significant interactions. Most importantly, Generalization × Block × F0 was significant, whereas Generalization × Block × F0 × Group was not [Generalization: $F(1, 25) = 4.552$, $p < .05$; F0: $F(1, 25) = 68.683$, $p < .001$; Generalization*Block*F0: $F(2, 50) = 18.388$, $p < .001$; Generalization*Block*F0*Group: $F(2, 50) = 2.121$, $p = .131$]. The main effect of Group was not significant, indicating that there was no overall group difference in the percent of voiceless responses, $F(1, 25) = 2.231$, $p = .148$. These results indicate that the influence of F0 on voicing judgments was modulated by block (baseline, canonical, reversed) and generalization type (exposed, generalization), and this interaction was consistent across the two groups (BP and DT groups). Post hoc tests were conducted to explore the Generalization × Block × F0 interaction.

Table 1
Factors of Group and Generalization in Experiment 2

	Group	
Generalization	BP exposure group	DT exposure group
Exposed words	*beer-pier*	*deer-tear*
Generalization words	*deer-tear*	*beer-pier*

Paired sample *t* tests examined the influence of F0 on voicing judgments (i.e., difference in percent voiceless response between high F0 and low F0) in each of the three blocks (baseline, canonical and reversed) across each generalization type (exposed and generalization) in the data collapsed across the two groups. The results showed that for exposed words, F0 affected voicing judgments in the baseline and the canonical correlation blocks [$p <$ .008 for both, alpha adjusted for 6 comparisons], but not in the reversed block [$p = .158$]. This means that F0 affected listeners' voicing judgments at the baseline, when there was no inherent F0/VOT correlation in the stimuli, and during the first canonical correlation block, when the F0/VOT correlation was compatible with the English norm. However, consistent with the findings of Idemaru and Holt (2011), listeners ceased to use F0 when they experienced a reversed F0/VOT correlation in short-term exposure.

For generalization words, the influence of F0 on voicing judgment was observed at baseline, as expected from long-term experience with English, and persisted through the exposure task across all blocks in a manner consistent with the canonical English correlation between F0 and VOT [$p <$ .008 for all]. An additional 3 × 2 (Block × F0) ANOVA for generalization words indicated a significant main effect of F0, whereas the main effect of Block and the Block × F0 interaction was not significant [F0: $F(1, 26) =$ 62.850, p = < .001; Block: $F(2, 52) = .184$, $p = .832$; Block*F0: $F(2, 52) = 2.122$, $p = .130$], indicating that the magnitude of F0 influence was equivalent across all blocks. This means that the

baseline influence of F0 in voicing judgments for generalization words was not modulated as a function of block and persisted through the experiment; experience with the reversed F0/VOT correlation at the other place of articulation did not influence listeners' use of F0 for generalization stimuli.

Discussion

Experiment 1 replicated our prior finding demonstrating that listeners track dimensional relationships in online speech processing to dynamically tune long-term speech representations to local regularities of the speech input (Idemaru & Holt, 2011). The diagnosticity of an acoustic dimension in signaling speech category membership is rapidly and dynamically adjusted in response to changing correlations between acoustic dimensions in short-term input. When the relationship of F0 to the perceptually unambiguous VOT cue signaling *beer* versus *pier* or *deer* versus *tear* category membership reversed in the local speech input, listeners down-weighted reliance on F0 in voicing decisions. In this dimension-based statistical learning, listeners' sensitivity to the relationship between acoustic dimensions defining a speech category (F0 and VOT) served to guide learning.

Most importantly, the current experiment also demonstrated that this dimension-based learning for one stop voicing contrast (e.g., *beer-pier*) did not generalize to the categorization of another stop voicing contrast (e.g., *deer-tear*). In other words, learning did not generalize to a new place of articulation (from bilabial to alveolar or vice versa) and did not extend, broadly, to the phonological category of voicing. Although listeners in the current experiment ceased to use F0 information in categorizing one contrast (e.g., *beer-pier*) for which they experienced a reversed F0/VOT correlation in short-term input, they simultaneously maintained canonical use of F0 in categorizing another contrast (e.g., *deer-tear*) within the same class of stop consonants. It is notable that these different patterns of perception were observed even though the two pairs of words were produced by the same talker. This finding suggests dimension-based statistical learning may be specific to correlations among acoustic dimensions across different phonetic categories even within the speech of a single talker, and furthermore, the system may be capable of tracking different statistical patterns across speech categories that are phonologically related (i.e., stop voicing).

In the current experiment, listeners experienced a shift of F0/VOT correlation in one contrast, but there were no local F0/VOT statistics available for the other contrast. As such, this was not the strongest test of whether listeners are able to track separate acoustic-dimension-based statistics across phonetic contrasts. Experiment 2 put this possibility to a rigorous test by exposing listeners to opposing F0/VOT statistics across two phonetic contrasts spoken by the same talker.

Experiment 2

Experiment 1 suggested the possibility that listeners can track dimension-based statistics across multiple phonetic categories that are typically classified together linguistically (i.e., stops). Experiment 2 investigates this possibility by presenting opposing statistics for the bilabial stop voicing contrast, *beer-pier,* and alveolar stop voicing contrast, *deer-tear,* spoken by the same talker. In a single block, listeners experienced the canonical F0/VOT relationship for one contrast and the reversed relationship for the other contrast. If listeners track the statistics of phonetic categories independently, then the predictions about reliance on F0 from prior research and Experiment 1 should be evidenced in opposing directions for the two contrasts. If, however, learning takes place generally at the level of stop class or for group statistics calculated across a block, then the canonical and reverse correlations should cancel one another and eliminate evidence of dimension-based statistical learning observed through down-weighting of F0 in voicing categorization.

Method

Participants. Thirty-three native-English listeners with normal hearing participated. They were either university students or employees. These participants were randomly assigned to Group 1 ($n = 15$) or Group 2 ($n = 18$).

Stimulus and procedure. The stimuli from Experiment 1 were used. As in Experiment 1, a categorization task measured the baseline influence of F0 in voicing judgment prior to the word recognition task. Two 5-ms steps were added to the 7-step VOT continua used in Experiment 1 to permit finer-grained assessment of the influence of F0 around the voicing boundary. The resulting continua included 9 VOT 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*). Each stimulus was presented 5 times, except for the middle VOT stimuli (step 5 of each series), which were presented 10 times so that the number of presentation of these critical stimuli was consistent across the baseline test and word recognition test. A total of 400 trials were randomized and presented in the baseline test, blocked for *beer-pier* and *deer-tear* types with the block order counterbalanced across participants.

In the subsequent word recognition task, the F0/VOT correlation characterizing the stimuli shifted in opposing directions for *beer-pier* and *deer-tear* stimuli across three exposure blocks (see Table 2). Listeners in Group 1 heard *beer* and *pier* with the canonical English F0/VOT correlation and *deer* and *tear* with the reversed F0/VOT correlation simultaneously in Block 1. The correlation then shifted to reversed for *beer* and *pier* and to canonical for *deer* and *tear* in Block 2, and then back to canonical for *beer* and *pier* and reversed for *deer* and *tear* in Block 3. This pattern was complementary for Group 2.

If listeners track the F0/VOT correlations separately for bilabial stops (i.e., *beer* and *pier*) and alveolar stops (i.e., *deer* and *tear*), we predict F0 down-weighting only for the stop pair for which the F0/VOT input correlation is reversed. The opposing correlations across place of articulation establishes no global F0/VOT correla-

Table 2
F0/VOT Correlation Patterns for Experiment 2

Group	Exposure word pair	F0/VOT correlation		
		Block 1	Block 2	Block 3
Group 1	*beer-pier*	Canonical	Reversed	Canonical
	deer-tear	Reversed	Canonical	Reversed
Group 2	*beer-pier*	Reversed	Canonical	Reversed
	deer-tear	Canonical	Reversed	Canonical

tion within a block at the level of voicing or, very generally, aggregate distributional statistics. If listeners track F0/VOT correlation at the level of voicing or aggregate distributional statistics, we expect no modulation of reliance upon F0 across blocks for either group.

The 20 exposure stimuli (open symbols, Figure 1) with perceptually unambiguous VOTs 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. There were a total of 600 exposure trials and 120 test trials. The apparatus and procedure were identical to Experiment 1, except that the visual icons presented on the monitor as response choices included all four pictures of a beer, a pier, a deer and a tear.

Results

Baseline test. A 2 × 9 × 2 × 2 (Place × VOT × F0 × Group) ANOVA with repeated measures on the first three factors was run on the mean percent voiceless responses to the baseline categorization test. The test revealed significant main effects of Place, VOT, F0, but not of Group [Place: $F(1, 31) = 43.212$, $p < .001$; VOT: $F(8, 248) = 971.127$, $p < .001$; F0: $F(1, 31) = 171.915$, $p < .001$; Group: $F(1, 31) = .065$, $p = .801$] as well as a number of significant interactions.

Critical to this experiment, a significant VOT × F0 interaction, $F(8, 248) = 44.467$, $p < .001$, indicated that the influence of F0 was modulated by VOT (Abramson & Lisker, 1985; Idemaru & Holt, 2011). A significant Place × VOT × F0 interaction and a marginally significant Place × VOT × F0 × Group interaction [Place*VOT*F0: $F(8, 248) = 6.508$, $p < .001$; Place*VOT*F0*Group: $F(8, 248) = 1.949$, $p = .054$] indicated that this modulation of the influence of F0 varied across *beer-pier* and *deer-tear* categorizations and that there was a trend toward a difference in this interaction between the two groups. To more closely examine this, the influence of F0 (i.e., the difference in percent voiceless response between high and low F0) was examined at the middle VOT value separately for each of *beer-pier* and *deer-tear* and for each group. Post hoc paired sample *t* tests revealed that F0 significantly affected categorization of both pairs for both groups [$p < .013$ for all, alpha adjusted for 4 comparisons]. The significant interactions are likely attributable to the difference in the magnitude of F0's influence across the pairs and groups (18.7% for *beer-pier* and 35.3% for *deer-tear* for Group 1; 13.3% for *beer-pier* and 21.1% for *deer-tear* for Group2). These results verified that listeners exhibited the canonical English influence of F0 on voicing judgments for both word pairs before the exposure experiment.

It is noted that at baseline Group 1 potentially was more influenced by F0, and both groups had a larger F0 effect for *deer-tear* categorization. The possible group difference will be noted in the interpretation of the results. The presence of stronger influence of F0 for *deer-tear* categorization was consistent with the results of multiple experiments from Idemaru and Holt (2011) and Experiment 1. It is also noted that the middle step-5 VOT value for *beer-pier* continua (10 ms) did not elicit the largest F0 effect for either group: the step-4 VOT (5 ms) elicited an average of 35.2% difference in percent voiceless response (49.3% by Group 1 and 23.3% by Group 2). This was not a critical issue as the focus of this experiment was whether the influence of F0 is modulated as a function of input statistics. However, we note that these particular stimuli may not have elicited the largest possible influence of F0 in *beer-pier* categorization.

Exposure 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 stop voicing. Because the block design was different across two word pairs and two groups, a separate 4 × 2 (Block × F0) repeated-measures 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* test indicated significant main effects of Block and F0, and a significant Block × F0 interaction [Block: $F(3, 42) = 2.872$, $p = .048$; F0: $F(1, 14) = 31.850$, $p < .001$; Block*F0: $F(3, 42) = 2.883$, $p < .05$]. Post hoc paired sample *t* tests indicated that F0 influenced *beer-pier* categorization at baseline, and in Blocks 1 and 3, when the F0/VOT correlation was canonical [$p < .013$, alpha adjusted for 4 comparisons], but not in the Block 2 when the F0/VOT correlation was reversed [$p = .191$]. The ANOVA for Group 1's *deer-tear* also indicated a significant main effect of F0 and a significant Block × F0 interaction [F0: $F(1, 14) = 32.367$, $p < .001$; Block*F0: $F(3, 42) = 4.694$, $p < .001$]. Post hoc paired sample *t* tests revealed that the F0 affected *deer-tear* categorization in the baseline and in Block 3 (canonical correlation; $p < .013$, alpha adjusted for 4 comparisons) but not in Blocks 1 and 2 (reversed correlation; $p = .020$ and .031). These results indicate that influence of F0 in voicing perception was modulated independently by exposure for the *beer-pier* contrast and the *deer-tear* contrast within each experimental block, paralleling the opposing patterns of F0/VOT correlation in the input listeners received.

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: $F(3, 51) = 3.113$, $p = .03$]. Post hoc paired sample *t* tests indicated that F0 influenced *beer-pier* categorization in the baseline ($p < .013$, alpha adjusted for 4 comparisons), but the influence disappeared in the remaining three blocks ($p = .399$, .076, and .730 for reversed, canonical and reversed correlation blocks). Thus, whereas the initial influence of F0 (baseline) disappeared when listeners encountered reversed F0/VOT correlation in the reversed block (Block 1), F0's influence did not bounce back when the F0/VOT correlation shifted to canonical (Block 2). In the canonical correlation block, there were on average more percent voiceless responses for the high F0 stimuli (78.3%) than for the low F0 stimuli (71.7%). Although the mean values were in the direction expected by exposure to the canonical F0/VOT correlation, the difference between the two F0 conditions did not reach statistical significance. Recall that there was a trend that the baseline influence of F0 was smaller for this group than the other group, and it is smaller for *beer-pier* categorization than in *deer-tear* categorization in the baseline test as well as in previous research (Idemaru & Holt, 2011; Experiment 1). It is possible that these factors contributed to the lack of an influence of F0 for this group's *beer-pier* categorization in the canonical correlation block.

The ANOVA for Group 2's *deer-tear* indicated a significant main effect of F0 and a significant interaction between Block and

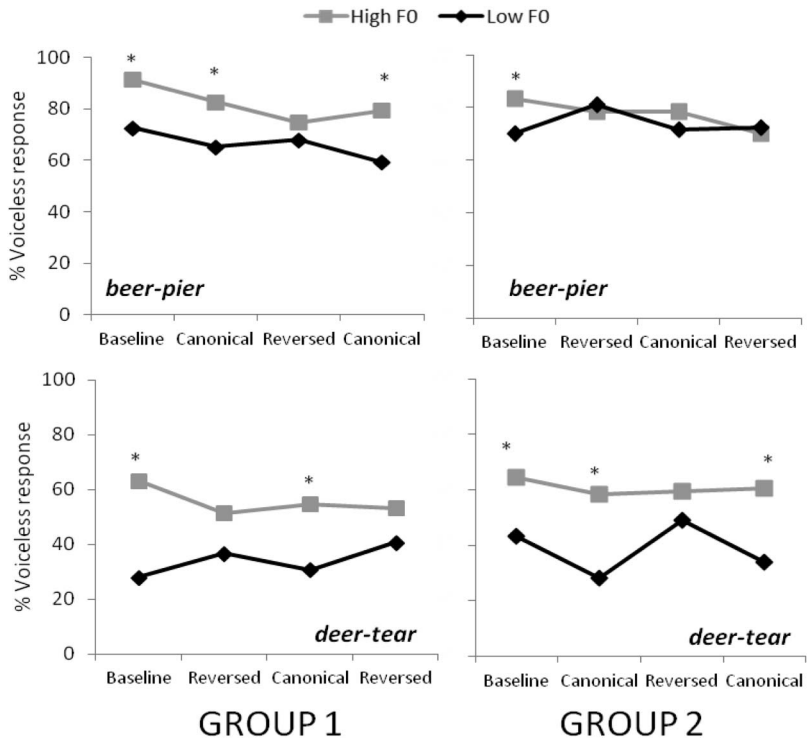

Figure 3. Mean percent voiceless response across four blocks for Group 1 and Group 2 in Experiment 2. A star indicates statistical significance ($p < .013$).

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 influenced *deer-tear* categorization in the baseline and in Blocks 1 and 3 (canonical correlation; $p < .013$, alpha adjusted for 4 comparisons), but F0 influence was absent in Block 2 (reversed correlation; $p = .103$). In this case, reliance upon F0 in informing voicing judgments was modulated as a function of F0/VOT correlations in short-term input.

Discussion

Listeners were not influenced by the global F0/VOT correlations within a block or across the experiment. If they were, the competing F0/VOT statistics across *beer-pier* and *deer-tear* would cancel one another, resulting in no acoustic F0/VOT correlation among exposure stimuli, and there would be no modulation of the influence of F0 in perception as a function of block. Instead, the results of this experiment indicate that listeners simultaneously track separate statistics defining voicing categories across different places of articulation.

In general, listeners ceased to rely on F0 in response to perceptually ambiguous VOT stimuli (e.g., ambiguous *beer-pier*) experienced with a reversed F0/VOT correlation, while simultaneously maintaining reliance on F0 in responding to perceptually ambiguous VOT stimuli at another place-of-articulation experienced with a canonical F0/VOT correlation (e.g., ambiguous *deer-tear*). When the F0/VOT correlation shifted in the course of the experiment in the opposing direction across the places of articulation, listeners' perceptual patterns shifted. Reliance on F0 in voicing judgments for the stops (e.g., *beer-pier*), for which F0 was previously down-weighted, returned when the F0/VOT correlation shifted back to the familiar long-term English pattern. At the same time, reliance on F0 in categorizing stops at the other place (e.g., *deer-tear*) was eliminated when the correlation reversed at this place of articulation.

Whereas Group 2's perception pattern did not show a perfect correspondence to the input F0/VOT statistics across the two word pairs as a result of listeners' failure to rely upon F0 for *beer-pier* word recognition in the context of canonical F0/VOT correlation, we nonetheless demonstrated that Group 2 listeners maintained separate statistics across *beer-pier* and *deer-tear* categorizations in Block 1 and Block 3. Whereas these listeners maintained the use of F0 in distinguishing alveolar stops, *deer* and *tear*, they simultaneously down-weighted use of F0 in distinguishing bilabial stops, *beer* and *pier*, in those blocks.

General Discussion

Speech categories are characterized by multiple acoustic dimensions, some of which carry more information in signaling category membership than others (Abramson & Lisker, 1985; Francis, Baldwin, & Nusbaum, 2000; Hillenbrand, Clark, & Houde, 2000; Idemaru, Holt, & Seltman, 2012). There is a long developmental course to acquiring native-like perceptual cue weightings, even among native listeners (e.g., Nittrouer, 1992; Hazan & Barrett, 2000; Idemaru & Holt, 2013). By adulthood, however, listeners exhibit reliable perceptual weights (Idemaru et al., 2012) that reflect regularities of acoustic dimensions in signaling native-

language speech categories (Lotto, Sato, & Diehl, 2004; Iverson et al., 2003; Yamada & Tohkura, 1992; Ingvalson, McClelland, & Holt, 2011; Escudero, Benders, & Lipski, 2009; Lipski, Escudero, & Benders, 2012; Francis, Kaganovich, & Driscoll-Huber, 2008). Perceptual cue weighting is a hallmark of mature phonetic categorization (Holt & Lotto, 2006; Toscano & McMurray, 2010).

Nevertheless, although perceptual cue weighting reflects long-term experience with distributional characteristics of native-language speech input, adult listeners also rapidly adjust perceptual weight in response to perturbations of long-term regularities. Idemaru and Holt (2011) showed that when the canonical English relationship of F0 to voicing categories, as expressed in *beer* versus *pier* and *deer* versus *tear*, changed in the local speech input, listeners adjusted the perceptual weight of F0 in judging the voicing contrast. Listeners rapidly down-weighted their reliance on F0 when the relationship of F0 to VOT reversed in short-term experience. The same listeners quickly returned to using F0 in voicing categorization when the short-term input returned to reflect the canonical F0/VOT relationship. This demonstrates *dimension-based statistical learning*, the rapid flexibility of perceptual cue weighting as a function of local input statistics at the level of fine-grained acoustic dimensions.

The present study investigated how dimension-based statistical learning generalizes as a means of examining the level at which learning occurs. The findings suggest a striking specificity of learning under the conditions we tested. In Experiment 1 the local short-term F0/VOT correlation reversed from the long-term English norm at a single place of articulation (e.g., bilabial stops, *beer* and *pier*). As such, the local input statistics lacked information about the F0/VOT relationship at the other place of articulation (e.g., alveolar stops, *deer* and *tear*). In this case, listeners down-weighted F0 in voicing categorization only for the stop place with which they experienced the short-term deviation; they maintained reliance upon F0 consistent with long-term perceptual experience for the other stop place. Thus, learning the relationship of F0/VOT for one place of articulation, /b-p/ or /d-t/, did not generalize to the other place of articulation, although both are classified as voicing contrasts. Said another way, the dimension-based statistical learning observed in Experiment 1 was specific to stop place category, not voicing. In Experiment 2, there were local statistics available for voicing at both bilabial and alveolar places of articulation (i.e., *beer* and *pier*, *deer* and *tear*) but the statistics were opposing such that globally, at the level of voicing, the correlation between F0 and VOT was neutral. In this case, listeners responded independently to each stop place category. Perceptual patterns mirrored the specific dimensional statistics of each place of articulation. Listeners down-weighted F0 in voicing categorization for the stop place for which the F0/VOT correlation was reversed (i.e., decreasing reliance on F0), but simultaneously maintained the long-term perceptual pattern (i.e., reliance on F0) for the place of articulation for which the F0/VOT correlation was consistent with the long-term English norm. Furthermore, when this pattern switched across places of articulation in the course of the experiment, listeners' perceptual pattern tracked the correlation change in the short-term distributional statistics such that they down-weighted F0 as a cue to voicing for the stop place of articulation that now had reversed F0/VOT signal statistics. The only exception for this was one group of listeners' (Group 2) responses. For this group, when the input statics shifted from reversed to the English norm among bilabial stops, *beer* and *pier*, perception did not shift back to reflect a reliance on F0. The same listeners, however, showed a perceptual switch when the statistics in alveolar stops, *deer* and *tear*, changed from reversed F0/VOT correlation to the English norm demonstrating that they, too, tracked the F0/VOT relationships independently across place of articulation.

Listeners are sensitive to input statistics defining stop voicing contrasts independently for bilabial and alveolar places of articulation (i.e., /b-p/ vs. /d-t/), even though the voicing pairs are from the same class of sounds (stop) and are typically contrasted together as voiceless (/p, t/) versus voiced stops (/b, d/). Thus, dimension-based statistical learning does not seem to operate at the level of stop class or stop voicing. Instead, it is specific to the details of experienced regularities of sounds within the class even for speech produced by the same talker. It is even possible that this learning may be specific to the experienced phonetic environment. Learning to down-weight the influence of F0 in *beer-pier* categorization may not generalize to *bear-pear* categorization, for example. It will be important for future research to determine the specificity of this learning and what constrains or enables generalization. Doing so will inform us about the structure of lower-level speech representations and their interaction.

In this context, it is worth noting that in our present and previous studies (Idemaru & Holt, 2011) the influence of F0 on voicing categorization, its perceptual weight, is consistently larger for the /d-t/ contrast than for the /b-p/ contrast at baseline. Moreover, although dimension-based statistical learning (as evidenced from the down-weighting of F0 as a result of a reversal of the F0/VOT correlation in short-term input) is reliably observed across both places of articulation, it is consistently less robust for /d-t/ than /b-p/. Overall, listeners' long-term perceptual weighting of F0 with respect to voicing appears to be more robust for alveolar stop consonants. We suggest that this pattern of perceptual data may inform predictions about the relationship of F0 and VOT in speech acoustics and the development of perceptual cue weighting across childhood. Given our results indicating that the perceptual representation of cue weights is independent across place of articulation, it is reasonable to posit that listeners may be sensitive to very fine-grained long-term differences in the informativeness of F0 to voicing categorization across place of articulation. Thus, we hypothesize that the correlation of F0 with VOT in English speech acoustics may be stronger at the alveolar, as compared to the bilabial, place of articulation. This would make F0 more informative for voicing categorization of alveolar (/d-t/) than bilabial (/b-p/) stops, consistent with listeners' greater reliance upon F0 in categorizing alveolar stop voicing at baseline and somewhat lesser malleability of F0 perceptual weight in response to deviations in short-term input. If such a difference in the correlation of F0 and VOT across places of articulation is found, it may be related to the fact that the acoustic range of VOT across voiced and voiceless categories is greater for alveolar stops, relative to bilabial stops (Lisker & Abramson, 1964). Perhaps the greater range sampled along the VOT dimension between voiceless, and voiced alveolar stops accentuates F0/VOT covariation more robustly than more restricted VOT range for bilabial stops. Furthermore, other acoustic dimensions defining the place of articulation (i.e., burst frequency and the formant transition in the following vowel) are more variable in alveolar than in bilabial stop production (e.g., Dorman, Studdert-Kennedy & Raphael, 1977; Liberman, Delattre, Cooper,

& Gerstman, 1954). It thus appears that the acoustic dimensions signaling alveolar stops are more variable and less tightly defined. One possibility is that greater acoustic variability in alveolar stop acoustics may be related to more robust reliance upon a secondary voicing cue, F0, at the alveolar place. Following the logic that the correlation of F0 with VOT in English is stronger at the alveolar place of articulation, we would expect developmental consequences of these patterns. If the correlation between F0 and VOT is more robust at the alveolar than the bilabial place of articulation, we expect that children will rely upon F0 in voicing categorization at an earlier age for alveolar, as opposed to bilabial, stop consonants. Thus, studying generalization patterns of short-term statistical learning relevant to speech categorization among adult listeners can inform hypotheses for speech production, perception, and development.

The finding that dimension-based statistical learning does not generalize across place of articulation and that listeners can track separate acoustic statistics signaling voicing across place of articulation also informs the details of linguistic representations. The observed specificity of this learning has implications for the relevance of the stop category for perceptual learning. There are synchronic phonological rules that apply across stop places of articulation in many languages (e.g., final obstruent devoicing). Thus, the level of stop class and that of stop voicing likely have linguistic reality for some processes. However, it is not clear whether the phonological category of stops or stop voicing is relevant in describing online perceptual processing such as dimension-based statistical learning. For this learning, the relevant level appears to be constrained to individual phonetic categories (e.g., /b/ vs. /p/ rather than voiced vs. voiceless stops). These implications regarding the independence of speech categories may be related to the idea of phonetically gradual sound change (e.g., Bybee, 2002). Phonetically gradual sound change describes the kind of change that affects production of a phonetic category (e.g., word-final /t/ and /d/ deletion, as in the production of *west,* in American English) earlier in some words than others instead of simultaneously affecting all words possessing the sound. Such sound change may be specific to a place of articulation (e.g., there is no word-final /p/ and /b/ deletion), and the occurrence of the change is affected by the frequency of the words that include these sounds. Namely, sounds in frequent words undergo change earlier relative to those of infrequent words. Such variation in speech production suggests category-specific representations of speech sounds and even variable representations of the same phonetic category in different lexical items.

The current results also have implications for understanding statistical learning. Investigation of statistical learning has typically focused on syllables, phonetic categories, or words as the functional units across which transitional probabilities, frequency-of-occurrence distributions, or nonadjacent dependency statistics are calculated (e.g., Saffran et al., 1996; Newport & Aslin, 2004). In natural speech, the acoustic information characterizing functional speech units like syllables or phonetic categories is itself probabilistic. Statistical regularity is particularly rich in speech because phonetic categories are inherently multidimensional at the level of acoustics and the acoustic dimensions that signal phonetic categories possess statistical regularity that is highly specific to native language and even dialect. Our results indicate that listeners exploit sensitivity to regularities at the level of acoustic dimensions to track and maintain multiple statistics across speech categories from short-term input. In natural speech input, where acoustics can be highly variable because of speaker, contextual, and idiosyncratic factors, the sensitivity to dimension-based statistics may be important in adapting perception to local acoustic regularities as they relate locally to speech categories. In spite of the considerable importance of dimension-based statistical learning as a potential mechanism for parsing variable acoustic speech signal, this type of statistical learning has not been studied extensively.

Dimension-based statistical learning occurs very rapidly. Our prior studies revealed that down-weighting of F0 in voicing judgment occurred already at the first presentation of test words after the F0/VOT correlation changed from the canonical to reversed (Idemaru & Holt, 2011). Recall that test words were interspersed among exposure words in all of our experiments. On average, just five instances each of *beer* or *pier* and *deer* or *tear* exposure words with reversed F0/VOT correlation were presented before the first test words (Idemaru & Holt, 2011). This indicates a highly responsive perceptual system adapting rapidly to the short-term regularities of incoming speech signals. Furthermore, the current findings have revealed that it is not only highly responsive, it is also very dynamic. In particular, the results of Experiment 2 reflect that listeners dynamically track two separate statistics. They alternated between using (long-term perceptual pattern) and not using F0 information (short-term perceptual pattern) in voicing judgments separately across place of articulation influenced by the input F0/VOT correlation. One may consider that listeners simply turned attention away from the F0 dimension when the F0/VOT correlation was reversed. However, the fact that short-term perceptual pattern returned when the F0/VOT correlation was later canonical indicates that listeners continued to track F0/VOT correlation as it changed through the course of the experiment. They adapted voicing judgments quickly, and independently, for /d-t/ and /b-p/ within the first few experiences of the new F0/VOT correlation, as suggested by our previous findings.

On what basis might listeners have identified the initial bilabial versus alveolar sounds in the stimuli to allow tracking of opposing statistics? The formant transitions for bilabial and alveolar stops followed by /**I**/ are similar across these places of articulation (Liberman et al., 1954; Delattre, Liberman, & Cooper, 1955), and this is verified in the acoustics of our stimuli. In this case, the most likely acoustic cue for place of articulation is the frequency of the noise burst at sound onset: it is lower-frequency for /b-p/ and higher for /d-t/ (Dorman et al., 1977). It is striking that listeners must have been sensitive to this subtle and minimal acoustic cue to differentiate stop places of articulation and differentially adjusted the perceptual weight given to F0 in voicing judgments when the informative F0 information occurred immediately after the burst and VOT (0–40 ms).

Many studies have found that lexical information can tune phonetic categorization such that unusual or otherwise distorted productions are subsequently more acceptable as category instances after being experienced in lexically biasing contexts (e.g., Norris et al., 2003; Eisner & McQueen, 2005; Kraljic & Samuel, 2006, 2007; Maye et al., 2002; Reinisch & Holt, 2013). Some of these studies evidence generalization across talker, at least when talkers are acoustically or perceptually similar (Eisner & McQueen, 2005; Kraljic & Samuel, 2006, 2007; Reinisch & Holt, 2013) or across words (Maye et al., 2002). The pattern of gener-

alization across segments observed for lexically guided phonetic tuning (Kraljic & Samuel, 2006) is particularly interesting in light of the present results. Whereas we find no evidence of generalization and, in fact, observe that listeners independently track statistics across place of articulation, Kraljic and Samuel (2006) report generalization of lexically guided phonetic tuning from /d-t/ to /b-p/.

Although lexically guided phonetic tuning studies and our study share the aim of understanding mechanisms by which speech processing adapts to accommodate short-term deviations from expectations driven by long-term input regularities such as those introduced by accent, there are several important differences. A notable difference is that lexical information (i.e., word vs. nonword) is not informative in the present paradigm. Lexically guided phonetic tuning is thought to be driven by feedback from lexical activation to influence sublexical representations (e.g., Norris et al., 2003; Mirman et al., 2006; McClelland et al., 2006). In the present paradigm, lexical information cannot provide an effective learning signal (e.g., *_ird* would give an effective learning signal to "teach" the system that the ambiguous stop should be heard as /b/, not /p/, because *bird* is a lexical item, whereas *pird* is not) because all stimuli and their voicing alternatives (*beer, pier, deer, tear*) are legitimate English words. Without lexical information to drive learning, the perceptually unambiguous VOT information that is consistently available to unambiguously signal word identity on all exposure trials may serve as the learning signal for adjusting the perceptual weight given to the correlated F0 input (Idemaru & Holt, 2011). The stimuli in the majority of trials in our paradigm are perceptually unambiguous with regard to the voicing category as signaled unequivocally by VOT (83.3%, 600 of 720 trials). Listeners accurately recognized these exposure stimuli 95% of the time in our previous studies (Idemaru & Holt, 2011). With VOT the primary voicing cue and F0 a secondary cue (e.g., Abramson & Lisker, 1985), F0 is never "required" of our word recognition task. Instead F0 is simply available in the signal as a reliable, correlated cue covarying with the critical VOT information. Our results therefore demonstrate bottom-up perceptual learning that can tune phonetic categories without necessitating top-down lexical feedback arising from lexically biasing context to drive learning. Although we chose to embed the task in the ecologically relevant task of word recognition, we expect this learning would occur even for nonword stimulus sets. Our results may implicate a reweighting of how bottom-up acoustic information influences activation of phonetic categories, with learning operating at a sublexical level. The generalization patterns of lexically guided phonetic tuning also implicate sublexical learning (Norris et al., 2003). Nonetheless, the different patterns of generalization across lexically guided tuning and dimension-based statistical learning suggest the possibility that top-down (lexical) versus bottom-up (acoustic dimension) learning signals may differentially impact phonetic processing. An important goal for future research is to understand more fully whether and how dimension-based statistical learning affects and interacts with lexical representation and, similarly, the extent to which lexically guided phonetic retuning depends upon the detailed distributions of acoustic information experienced during learning.

The current paradigm is similar to Clayards et al. (2008) in the very point that perceptual learning investigated does not require lexical feedback. In Clayards et al. (2008) the frequency distribution of VOT values in the stimuli served as a teaching signal to tune subsequent phonetic categorization. One stimulus set had a narrower range of VOTs characterizing voiced and voiceless stops in *beach-peach, beak-peak,* and *bees-peas* and the category center value of VOT occurred more frequently. Thus, as in our studies, lexical information did not bias interpretation, as all phonetic possibilities formed real words. Another stimulus set composed of the same words sampled a wider range of VOTs and the center value of VOT occurred less frequently. Exposure to these different distributions sampling the VOT dimension led listeners to show different levels of certainty in subsequent voicing judgments, with steeper categorization functions indicating more distinct category judgments after exposure to the narrow distributions. Thus listeners appear to be sensitive to frequency-of-occurrence distributional statistics, at least along a single acoustic dimension. In our previous (Idemaru & Holt, 2011) and present work, listeners exhibit learning across two acoustic dimensions signaling phonetic categories (VOT and F0) and their covariation (the correlation of F0 with VOT). Because natural spoken language is likely to vary in both types of distributional information, it will be important to understand the relationship of these two effects of statistical information on phonetic tuning. Clayards et al. (2008) did not test generalization and acknowledge that the learning they observe could take place at lexical, phonetic, syllabic, or featural levels. Examining generalization of learning for frequency-of-occurrence statistics as in Clayards et al. (2008) and its relationship to the present findings will be informative. There remain important theoretical and empirical questions to determine the nature of these effects and to develop detailed models that predict patterns of generalization.

Ultimately, using generalization as a means of understanding the extent to which common mechanisms contribute to phonetic tuning across different information sources will require convergence in experimental approaches as task differences are a possible source of differences in patterns of generalization. Whereas lexically guided phonetic tuning studies typically use 20 words with the accented pronunciation (e.g., Kraljic & Samuel, 2006), our paradigm had just four words (i.e., *beer, pier, deer* and *tear*), with only two accented words experienced in Experiment 1. Prior research has demonstrated that details of stimulus-list construction can influence the strength with which information sources contribute to learning (Mirman et al., 2008; Pitt & Szostak, 2012; Reinisch & Holt, submitted). It is possible that the number of different words experienced with the accented pronunciation of a phoneme may be a critical factor in generalization. Use of multiple word pairs (i.e., *beach-peach, beak-peak,* and *bees-peas*) in the stimulus list may have encouraged listeners to track the VOT distribution in these words more robustly in Clayards et al. (2008) than they would have had they encountered a single word pair, for example. Listeners in our task experienced information for a single word pair from a single talker at each place of articulation. In this context, our paradigm presents a very conservative experimental test of the information needed to evoke dimension-based statistical learning. Examinations under conservative conditions are important in understanding the amount of speech data that is necessary to engage perceptual learning. It is possible that testing in situations closer to natural listening conditions—for example, using larger word lists or continuous speech—may encourage greater generalization. Investigating the conditions under which general-

ization does and does not occur will be important in understanding the nature of the learning and, potentially, in differentiating learning arising from different information sources. Future experiments might tease apart the relative influence of bottom-up statistical regularities and top-down lexical information in perceptual learning at a sublexical level by manipulating stimulus lists to encourage greater or lesser reliance on lexical versus statistical information (see Reinisch & Holt, submitted for an example). Even with these caveats, the present data demonstrate unequivocally that the system is not bound necessarily by phonological voicing; perceptual cue weighting can be influenced by short-term deviations in correlations between acoustic dimensions at the level of phonetic categories and independent statistics can be tracked across place of articulation even for categories within a common phonological class and for tokens spoken by a single talker.

Studies on perceptual learning and phonetic tuning have demonstrated that perceptual learning can be induced by various teaching signals, including lexical (Norris et al., 2003; Eisner & McQueen, 2005; Kraljic & Samuel, 2006, 2007; Maye et al., 2002; Reinisch & Holt, 2013), visual (Bertelson et al., 2003; Vroomen et al., 2007), phonotactic (Cutler et al., 2008), and statistical information (Idemaru & Holt, 2011; Clayards et al., 2008). Examining how details of various learning signals differentially impact short-term adaptation to distorted or ambiguous speech input provides information with which to constrain models of speech representations and how these representations interact. Our findings suggest that at the sublexical level, at which we presume dimension-based statistical learning occurs, speech categories within the same phonological voicing class are represented independently.

References

Abramson, A. S., & Lisker, L. (1985). Relative power of cues: F0 shift versus voice timing. In V. Fromkin (Ed.), *Phonetic linguistics: Essays in honor of Peter Ladefoged* (pp. 25–33). New York, NY: Academic.

Bertelson, P., Vroomen, J., & de Gelder, B. (2003). Visual recalibration of auditory speech identification: A McGurk aftereffect. *Psychological Science, 14,* 592–597. doi:10.1046/j.0956-7976.2003.psci_1470.x

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

Bybee, J. (2002). Word frequency and context of use in the lexical diffusion of phonetically conditioned sound change. *Language Variation and Change, 14,* 261–290. doi:10.1017/S0954394502143018

Clayards, M., Tanenhaus, M. K., Aslin, R. N., & Jacobs, R. A. (2008). Perception of speech reflects optimal use of probabilistic speech cues. *Cognition, 108,* 804–809. doi:10.1016/j.cognition.2008.04.004

Cutler, A., McQueen, J. M., Butterfield, S., & Norris, D. (2008). Prelexically-driven perceptual retuning of phoneme boundaries. In J. Fletcher, D. Loakes, M. Wagner, & R. Goecke. (Eds.), *Proceedings of Interspeech 2008* (2056). Brisbane, Australia: ISCA.

Delattre, P. C., Liberman, A. M., & Cooper, F. S. (1955). Acoustic loci and transitional cues for consonants. *Journal of the Acoustical Society of America, 27,* 769. doi:10.1121/1.1908024

Dorman, M. F., Studdert-Kennedy, M., & Raphael, L. J. (1977). Stop-consonant recognition: Release bursts and formant transitions as functionally equivalent, context-dependent cues. *Perception & Psychophysics, 22,* 109–122. doi:10.3758/BF03198744

Eisner, F., & McQueen, J. M. (2005). The specificity of perceptual learning in speech processing. *Perception & Psychophysics, 67,* 224–238. doi:10.3758/BF03206487

Escudero, P., Benders, T., & Lipski, S. C. (2009). Native, non-native and L2 perceptual cue weighting for Dutch vowels: The case of Dutch, German, and Spanish listeners. *Journal of Phonetics, 37,* 452–465. doi:10.1016/j.wocn.2009.07.006

Francis, A. L., Baldwin, K., & Nusbaum, H. C. (2000). Effects of training on attention to acoustic cues. *Perception & Psychophysics, 62,* 1668–1680. doi:10.3758/BF03212164

Francis, A. L., Kaganovich, N., & Driscoll-Huber, C. (2008). Cue-specific effects of categorization training on the relative weighting of acoustic cues to consonant voicing in English. *Journal of the Acoustical Society of America, 124,* 1234–1251. doi:10.1121/1.2945161

Haggard, M., Ambler, S., & Callow, M. (1970). Pitch as a voicing cue. *Journal of the Acoustical Society of America, 47,* 613–617. doi:10.1121/1.1911936

Hazan, V., & Barrett, S. (2000). The development of phonemic categorization in children aged 6–12. *Journal of Phonetics, 28,* 377–396. doi:10.1006/jpho.2000.0121

Hillenbrand, J. M., Clark, M. J., & Houde, R. A. (2000). Some effects of duration on vowel recognition. *Journal of the Acoustical Society of America, 108,* 3013–3022. doi:10.1121/1.1323463

Holt, L. L., & Lotto, A. J. (2006). Cue weighting in auditory categorization: Implications for first and second language acquisition. *Journal of the Acoustical Society of America, 119,* 3059–3071. doi:10.1121/1.2188377

Idemaru, K., & Holt, L. L. (2011). Word recognition reflects dimension-based statistical learning. *Journal of Experimental Psychology: Human Perception and Performance, 37,* 1939–1956. doi:10.1037/a0025641

Idemaru, K., & Holt, L. L. (2013). The developmental trajectory of children's perception and production of English/r/-/l. *Journal of the Acoustical Society of America, 133,* 4232–4246. doi:10.1121/1.4802905

Idemaru, K., Holt, L. L., & Seltman, H. (2012). Individual differences in cue weights are stable across time: The case of Japanese stops lengths. *Journal of the Acoustical Society of America, 132,* 3950–3964. doi:10.1121/1.4765076

Ingvalson, E. M., McClelland, J. M., & Holt, L. L. (2011). Predicting native English-like performance by native Japanese speakers. *Journal of Phonetics, 39,* 571–584. doi:10.1016/j.wocn.2011.03.003

Iverson, P., Kuhl, P. K., Akahane-Yamada, R., Diesch, E., Tohkura, Y., Kettermann, A., & Siebert, C. (2003). A perceptual interference account of acquisition difficulties for non-native phonemes. *Cognition, 87,* 47–57. doi:10.1016/S0010-0277(02)00198-1

Kraljic, T., & Samuel, A. G. (2006). Generalization in perceptual learning for speech. *Psychonomic Bulletin & Review, 13,* 262–268. doi:10.3758/BF03193841

Kraljic, T., & Samuel, A. G. (2007). Perceptual adjustments to multiple speakers. *Journal of Memory and Language, 56,* 1–15. doi:10.1016/j.jml.2006.07.010

Liberman, A. M., Delattre, P. C., Cooper, F. S., & Gerstman, L. J. (1954). The role of consonant-vowel transitions in the perception of the stop and nasal consonants. *Psychological Monographs: General and Applied, 68,* 1–13. doi:10.1037/h0093673

Lipski, S. C., Escudero, P., & Benders, T. (2012)., Language experience modulates weighting of acoustic cues for vowel perception: An event-related potential study. *Psychophysiology, 49,* 638–650. doi:10.1111/j.1469-8986.2011.01347.x

Lisker, L., & Abramson, A. S. (1964). A cross-language study of voicing in initial stops: Acoustical measurements. *Word, 20,* 384–422.

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 J. Slifka, S. Manuel, & M. Matthies. (Eds.), *From sound to sense: 50+ years of discoveries in speech communication* (pp. 181–186). Electronic conference proceedings.

Maye, J., Werker, J. F., & Gerken, L. A. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition, 82,* 101–111. doi:10.1016/S0010-0277(01)00157-3

McClelland, J. L. Mirman, D., & Holt, L. L. (2006). Are there interactive processes in speech perception? *Trends in Cognitive Sciences, 10,* 363–369. doi:10.1016/j.tics.2006.06.007

Mirman, D., McClelland, J. L., & Holt, L. L. (2006). An interactive Hebbian account of lexically guided tuning of speech perception. *Psychonomic Bulletin & Review, 13,* 958–965. doi:10.3758/BF03213909

Mirman, D., McClelland, J. L., Holt, L. L., & Magnuson, J. S. (2008). Effects of attention on the strength of lexical influences on speech perception: Behavioral experiments and computational mechanisms. *Cognitive Science, 32,* 398–417. doi:10.1080/03640210701864063

Newport, E. L., & Aslin, R. N. (2004). Learning at a distance I. Statistical learning of non-adjacent dependencies. *Cognitive Psychology, 48,* 127–162. doi:10.1016/S0010-0285(03)00128-2

Nittrouer, S. (1992). Age-related differences in perceptual effect of formant transitions within syllables and across syllable boundaries. *Journal of Phonetics, 20,* 351–382.

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology, 47,* 204–238. doi:10.1016/S0010-0285(03)00006-9

Pitt, M. A., & Szostak, C. M. (2012). A lexically biased attentional set compensates for variable speech quality caused by pronunciation variation. *Language and Cognitive Processes, 27,* 1225–1239. doi:10.1080/01690965.2011.619370

Reinisch, E., & Holt, L. L. (manuscript submitted for publication). *Listening situation modulates lexical and acoustic context effects in phonetic categorization.*

Reinisch, E. & Holt, L. L. (2013). Lexically-guided phonetic retuning of foreign-accented speech and its generalization. *Journal of Experimental Psychology: Human Perception and Performance.*

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science, 274,* 1926–1928. doi:10.1126/science.274.5294.1926

Toscano, J. C., & McMurray, B. (2010). Cue integration with categories: Weighting acoustic cues in speech using unsupervised learning and distributional statistics. *Cognitive Science, 34,* 434–464. doi:10.1111/j.1551-6709.2009.01077.x

Vroomen, J., van Linden, S., de Gelder, B., & Bertelson, P. (2007). Visual recalibration and selective adaptation in auditory-visual speech perception: Contrasting build-up courses. *Neuropsychologia, 45,* 572–577. doi:10.1016/j.neuropsychologia.2006.01.031

Whalen, D. H., Abramson, A. S., Lisker, L., & Mody, M. (1993). F0 gives voicing information even with unambiguous voice onset times. *Journal of the Acoustical Society of America, 93,* 2152–2159. doi:10.1121/1.406678

Yamada, R. A., & Tohkura, Y. (1992). The effects of experimental variables on the perception of American English /r/ and / l / by Japanese listeners. *Perception & Psychophysics, 52,* 376–392. doi:10.3758/BF03206698

Received June 7, 2013
Revision received September 16, 2013
Accepted October 22, 2013 ■