

# Word Recognition Reflects Dimension-based Statistical Learning

Kaori Idemaru  
University of Oregon

Lori L. Holt  
Carnegie Mellon University

Speech processing requires sensitivity to long-term regularities of the native language yet demands listeners to flexibly adapt to perturbations that arise from talker idiosyncrasies such as nonnative accent. The present experiments investigate whether listeners exhibit *dimension-based statistical learning* of correlations between acoustic dimensions defining perceptual space for a given speech segment. While engaged in a word recognition task guided by a perceptually unambiguous voice-onset time (VOT) acoustics to signal *beer*, *pier*, *deer*, or *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) to VOT. Results across four experiments are indicative of rapid, dimension-based statistical learning; reliance on the F0 dimension in word recognition was rapidly down-weighted in response to the perturbation of the correlation between F0 and VOT dimensions. However, listeners did not simply mirror the short-term input statistics. Instead, response patterns were consistent with a lingering influence of sensitivity to the long-term regularities of English. This suggests that the very acoustic dimensions defining perceptual space are not fixed and, rather, are dynamically and rapidly adjusted to the idiosyncrasies of local experience, such as might arise from nonnative-accent, dialect, or dysarthria. The current findings extend demonstrations of “object-based” statistical learning across speech segments to include incidental, online statistical learning of regularities residing *within* a speech segment.

*Keywords:* speech perception, perceptual learning, statistical learning, talker adaptation, dimension-based learning

The acoustics of speech are highly variable. Yet rich regularities reside within the variability, and accumulating evidence indicates that listeners make use of regularity in parsing the acoustic speech signal. Listeners are sensitive to transitional probabilities across syllables such that after just two minutes of exposure to a novel stream of nonsense syllables they detect that some syllables co-occur more consistently than others (Saffran, Aslin & Newport, 1996). Nonadjacent dependencies across units of speech are also detected by listeners (Newport & Aslin, 2004), and it appears that the frequency (Maye, Werker & Gerken, 2002) and variability (Clayards, Tanenhaus, Aslin & Jacobs, 2008) with which specific speech exemplars occur shape subsequent speech perception. Statistical learning of this sort demonstrates that listeners extract much regularity from speech, providing a means of perceptually organizing incoming spoken language.

Thus far, investigations of statistical learning of speech typically have focused, implicitly or explicitly, on what might be considered to be the “object” level whereby familiar syllables, phonetic categories, or words serve as the functional units across which tran-

sitional probabilities, frequency-of-occurrence distributions, or nonadjacent dependency statistics are calculated. The functional units, or objects, of statistical learning are defined a priori in these studies, tend to be drawn from a closed set, and tend to be acoustically invariant and familiar to the learners. In natural speech the acoustic information that characterizes functional speech units like syllables or phonetic categories is itself probabilistic; it thus presents its own learning challenge. Thus, we distinguish between “object-based” learning whereby relationships (like transitional probabilities) among functional units are learned and what might be considered to be “feature-“, “cue-,” or “dimension-based” learning for which the level of regularity to be learned resides *within* the functional units (Turk-Browne, Isola, Scholl, & Treat, 2008 for an example from visual learning). In the present work, we focus specifically on regularities that occur among the acoustic dimensions that define phonetic categories.

Dimension-based regularity is particularly rich in speech because phonetic categories are inherently probabilistic and multidimensional. Typically, no single acoustic dimension is necessary or sufficient to define phonetic category membership. Acoustic dimensions thus covary in the input (Coleman, 2003; Dorman, Studdert-Kennedy, & Raphael, 1977 for stop place of articulation; Jongman, Wayland, & Wong, 2000 for fricative place of articulation; Hillenbrand, Clark, & Houde, 2000 for tense and lax vowels; Kluender, & Walsh, 1992 for fricative/affricate distinction; Lisker, 1986 for stops voicing; Polka & Strange, 1985 for liquids) and differ in the degree to which they correlate with phonetic categories. Some acoustic dimensions are more diagnostic of category membership (they better correlate with category identity) and may be perceptually weighted more than other dimensions (Francis et

---

This article was published Online First October 17, 2011.

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

We thank Christi Gomez for running the experiments. This research was supported by the National Institutes of Health (R01DC004674), the National Science Foundation (0746067), and National Organization for Hearing Research.

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

al., 2008; Holt & Lotto, 2006; Iverson & Kuhl, 1995; Nittrouer, 2004).

A long history of empirical research exists demonstrating that listeners are sensitive to these relationships within the native language (Dorman, Studdert-Kennedy, & Raphael, 1977; Hillenbrand, Clark, & Houde, 2000; Kluender, & Walsh, 1992; Whalen, Abramson, Lisker & Mody, 1993). For example, perception of voicing (as in the difference between *beer* vs. *pier*) can be influenced by as many as 16 acoustic dimensions that distinguish [b] from [p] (Lisker, 1986). Whereas any of these multiple dimensions may inform categorization, their perceptual effectiveness varies. In the case of voiced consonants [b], [d], and [g], for example, American English listeners make greater use of differences in formant transitions than frequency information in the noise burst that precedes the transitions although each reliably covaries with these consonant categories (Francis et al., 2000). Moreover, listeners' relative reliance on particular acoustic dimensions changes across development (e.g., Mayo & Turk, 2005; Nittrouer, 2004; Narayan, to appear) and varies depending on native language (e.g., Iverson et al., 2003). Thus, speech categories are defined by multiple, simultaneous probabilistic dimensions that covary with one another and vary in their perceptual weighting.

There is strong evidence that listeners make use of this redundant information in speech categorization. For example, in English and many other languages, fundamental frequency (F0, related to the pitch of a voice) varies with voicing categories such that voiced consonants like [b] and [d] are produced with lower fundamental frequencies (F0s) than are voiceless consonants like [p] and [t] (Kingston & Diehl, 1994; Kohler, 1982; Kohler, 1984; Kohler, 1985). The regularity of F0 with voicing has most often been investigated with voice onset time (VOT) as an acoustic dimension of voicing. Figure 1 illustrates distributions of speech production

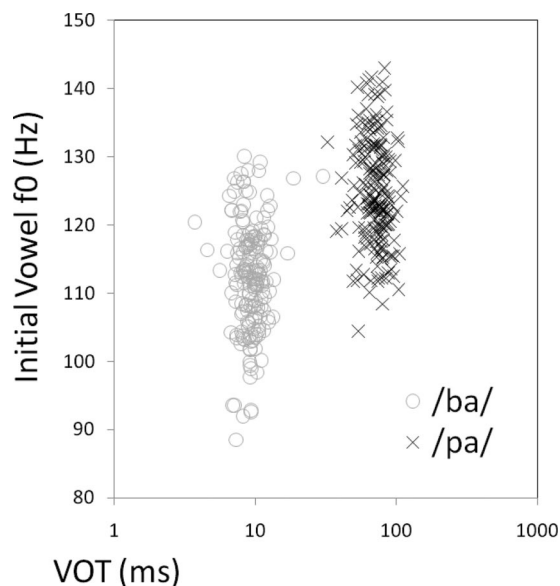


Figure 1. Fundamental frequency (F0 of the following vowel, in Hz) and voice onset time (VOT, in ms) are plotted for 400 utterances of syllable-initial [b] and [p] by a single male talker. Note the correlation between F0 and VOT such that voiceless [p], with longer VOT, tends to be produced with relatively higher F0 frequencies.

measurements of F0 and VOT across 400 unique productions of words beginning with [b] or [p] from a single voice (Holt & Wade, 2004); although the distributions overlap considerably, the voiceless consonant [p] has a higher F0 than [b]. This long-term regularity robustly influences speech perception such that syllables with an ambiguous VOT are more often perceived as beginning with [b] when synthesized or spoken with a subsequent vowel with a lower F0 and more often as [p] with a vowel with a higher F0 (Abramson & Lisker, 1985; Castleman & Diehl, 1996; Chistovich, 1969; Haggard, Ambler, & Callow, 1970; Haggard, Summerfield, & Roberts, 1981; Whalen, Abramson, Lisker, & Mody, 1993).

Listeners are thus sensitive to long-term experience with dimension-based regularities within speech categories. However, as Figure 1 makes clear, the correlation between dimensions is far from perfect even for a single talker uttering speech in a laboratory setting. There is even more variability in the realization of this regularity in natural speech; the F0 differences between male and female voices, alone, shift the experienced distributions considerably in acoustic space. Overall, there are distributional regularities in how acoustic dimensions relate to one another in defining multidimensional phonetic categories, but how a particular acoustic dimension is used in relation to another dimension varies quite a lot in speech input as a function of dialect, accent, and speaker idiosyncrasies.

In the present work, we investigate whether, in addition to sensitivity to long-term correlations among dimensions defining phonetic categories, listeners also use short-term regularities characterizing the present speech input to guide online perception. Specifically, we hypothesize that listeners may exhibit rapid *dimension-based statistical learning* such that speech categorization is influenced by the diagnosticity of an acoustic dimension in signaling category identity as it varies in the short term. Such rapid dimension-based statistical learning may be important for speech recognition because a source of variability in speech acoustics arises from talkers' linguistic experience, producing non-native accent, dialect, or idiosyncratic patterns of speech. A consequence of these variable patterns of speech production for listeners is that phonetic categories founded on long-term experience across a variety of talkers may produce mis-categorization of speech of a *particular* idiosyncratic (e.g., foreign-accented) talker. The long-term regularities of speech input to which listeners' perception is tuned may not adequately capture the regularity in a particular instance of speech in the short term. As an example, native English speakers learning Korean use the canonical English relationship of VOT and F0 (higher F0 for voiceless consonants) when producing Korean consonants, even though this relationship is not characteristic of Korean (Kim & Lotto, 2002). This non-native instantiation of Korean thus violates typical correlations among dimensions defining native Korean speech categories and presents a perceptual challenge for native Korean listeners.

Building upon this example, we investigate whether listeners exhibit dimension-based learning by exploiting the natural long-term correlation of F0 and voicing (as manipulated acoustically via changes in VOT) in English. Rather than investigating non-native speech, we use resynthesis techniques to manipulate a single talker's natural utterances such that the ordinary relationship between F0 and voicing is reversed in the course of the experiment. Thus, characteristics such as vocal tract shape and subtle acoustic cues distinctive to the talker are preserved whereas the direction of

the F0/voicing correlation reverses to opposite that of natural English productions. In essence, we confer a very strong “accent” upon this voice that violates a robust regularity of English. This is an extreme case of accent, to be sure; nevertheless, it is not without precedent. Acoustic measurements of native Japanese-accented English speech productions indicate that some Japanese-accented talkers reverse the relationship of second (F2) and third (F3) formant frequencies characteristic of American-accented English [r] and [l] (Lotto et al., 2004). Also, as noted, native-English speakers learning the Korean three-way stop consonant distinction fail to produce the correct Korean F0/voicing relationship (Kim & Lotto, 2002). Our method provides a means of rigorously controlling characteristics of acoustic dimensions to investigate how perturbations in listeners’ short-term experience impact categorization while using naturalistic materials created from and closely resembling natural speech.

Across four experiments, native-English listeners heard artificially “accented” words (rhymes *beer*, *pier*, *deer*, and *tear*) with F0/VOT correlations that deviated from typical English experience. On most trials, VOT, the primary acoustic cue for voicing (Francis et al., 2008), unambiguously signaled the voicing category, providing the opportunity to manipulate listeners’ experience with the correlation of VOT with F0. Our hypothesis is that as listeners recognize these “accented” words using the robust VOT cue, they also will track changes in the distributional patterns of F0 that covary with VOT. We hypothesize that online sensitivity to the relationship between F0 and VOT will affect how listeners respond to test stimuli with high or low F0 and ambiguous VOT values surreptitiously interspersed throughout the exposure trials. This approach allows us to examine the influence of changes in the correlation of acoustic dimensions *within* a speech segment on online word recognition.

### Experiment 1

In Experiment 1, listeners heard words (rhymes *beer*, *pier*, *deer*, and *tear*) and responded by clicking on a picture that matched the word. Unbeknownst to listeners, the relationship between F0 and VOT of the initial syllables presented as the exposure stimuli shifted across three implicit experimental blocks, from the familiar canonical English correlation (Canonical correlation, Block 1), to no correlation between VOT and F0 (Neutral correlation, Block 2),

to the “accented,” reversed correlation opposite long-term English experience (Reversed correlation, Block 3). A subset of words in the task (again, rhymes *beer*, *pier*, *deer*, and *tear*) was composed of test stimuli with ambiguous VOT values and either High or Low frequency F0s. Test trials were surreptitiously interspersed among the exposure trials throughout the experiment to assess listeners’ use of F0 in word recognition.

### Method

**Participants.** Fourteen native-English listeners participated for university credit or a small payment. They were either university students or employees. All reported normal hearing.

**Exposure stimuli.** Participants responded to exposure words as *beer*, *pier*, *deer*, or *tear* with perceptually *unambiguous* VOT values differentiating the voicing categories. The purpose of these stimuli was to manipulate participants’ short-term experience with the relationship between F0 and VOT. Figure 2 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.

In Block 1, listeners heard speech with the familiar, canonical English F0/VOT relationship (Canonical correlation): voiced stops in *beer* and *deer* had lower F0s whereas voiceless stops in rhymes *pier* and *tear* had higher F0s. Three perceptually unambiguous short VOT values (e.g., -20, -10, and 0 ms for the *beer/pier* series, heard as [b]) were combined with three low F0s (220, 230, and 240 Hz) whereas three long VOT values (e.g., 20, 30, and 40 ms for the *beer/pier* series, heard as [p]) were combined with three high F0s (280, 290 and 300 Hz). In Block 2, F0 did not correlate with VOT and was not diagnostic of voicing categories (Neutral correlation). Stimuli had one of three mid-F0 values (250, 260, or 270 Hz). In Block 3, the F0/VOT correlation was reversed such that listeners heard productions of *beer*, *pier*, *deer*, and *tear*, with an F0/VOT correlation opposite their long-term experience with English; high F0s (280, 290, and 300 Hz) were now paired with voiced consonants whereas low F0s (220, 230, and 240 Hz) were associated with voiceless sounds (Reversed correlation).

Thus, across blocks the correlation of F0 and VOT shifted from the natural pattern that characterizes the long-term regularities of English to a correlation pattern opposite that of English. In each block, each exposure stimulus was presented 10 times in a random

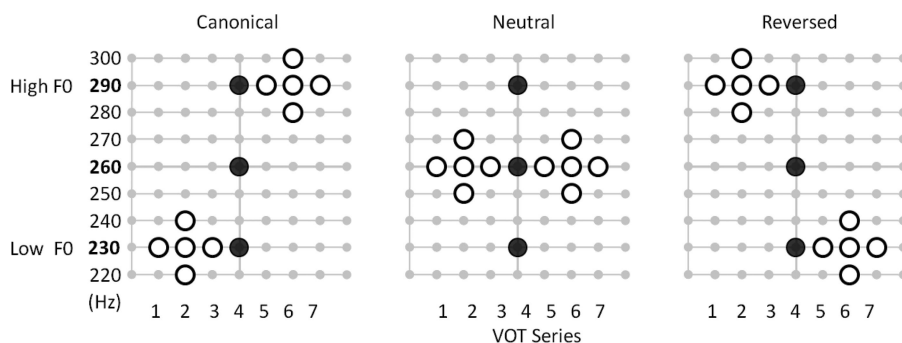


Figure 2. Schematic illustration of stimulus distributions across experiment blocks, defined by the VOT dimension in stimulus step (horizontal axis, see text for VOT values in ms) and F0 dimension (vertical axis, in Hz). Clear dots were exposure stimuli, and filled dots were critical test stimuli.

order. The block structure was implicit, serving to define the type of stimuli presented, but it was not apparent in the nature of the task. Participants were not informed that the experiment was divided into blocks or that the characteristics of the stimuli would vary. 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 filled symbols in Figure 2). F0 exerts the strongest influence on voicing perception when VOT is ambiguous (Abramson & Lisker, 1985; Francis et al., 2008) 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 correlations between F0 and VOT.

The test stimuli were constant across blocks and possessed perceptually ambiguous VOT values (10 ms for the *beer/pier* series and 20 ms for the *deer/tear* series, accommodating the shift in VOT category boundary that occurs across place-of-articulation; Abramson & Lisker, 1985) with low-, mid-, and high-F0 frequencies (230, 260, and 290 Hz) corresponding to the midpoint F0 frequencies of the exposure stimuli within these ranges. The six test stimuli (*beer/pier*, *deer/tear* × 3 F0s) and 20 exposure stimuli formed a unique list of stimuli. This list of stimuli was presented in 10 different randomized orders per block, with different randomization patterns for each listener. Thus, each test stimulus was presented 10 times per block, interspersed randomly among the exposure stimuli. The test stimuli were not described to participants, and they were not differentiated from exposure stimuli by task.

Based on the natural F0/VOT correlation in English (e.g., Abramson & Lisker, 1985) and previous perceptual results (e.g., Haggard, Ambler & Callow, 1970), we expected that native-English listeners would identify the low-F0 test stimuli more often as voiced (*beer* and *deer*) and the high-F0 test stimuli as voiceless (*pier* and *tear*) at the outset of the experiment. The extent to which this pattern of perception changes with manipulation of the experienced short-term F0/VOT correlation provides a measure of listeners' online sensitivity to dimension-based regularity within a speech segment.

**Stimulus creation.** Natural utterances of *beer*, *pier*, *deer*, and *tear* ([bɪər], [pɪər], [dɪər], and [tɪər]) were digitally recorded (22.05 kHz sampling frequency) from utterances of a female monolingual English native speaker (LLH). Words were spoken in isolation in citation form. Using these utterances as end points, voice onset time (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 pilot categorization tests indicating a category boundary at about 10-ms VOT for *beer/pier* series and 20-ms VOT for *deer/tear* for this speaker. The shift in voicing category boundary with place-of-articulation is typical of English voicing perception (Abramson & Lisker, 1985).

An instance of *pier* and an instance of *tear* were chosen based on clarity and their roughly equivalent duration. The waveforms of these voiceless end points were edited (see Francis et al., 2008) to create two series: one that varied perceptually from *beer* to *pier*

and another from *deer* to *tear*. The first 10 ms of the original voiceless productions were left intact to preserve the consonant bursts. Manipulation of VOT across the series was accomplished by removing approximately 10-ms segments (with minor variability so that edits were made at zero-crossings) from the waveform using Praat 5.0 (Boersma & Weenink, 2010). 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) of the two series was manipulated such that the F0 onset frequency of the vowel, [ɪ], following the word-initial stop consonant was adjusted from 220 Hz to 300 Hz across nine 10-Hz steps. For each stimulus, the F0 contour of the original production was measured and manually manipulated using Praat 5.0 to adjust the target onset F0. As Figure 3 shows, 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.

**Procedure.** Participants were seated in front of a computer monitor in a quiet room. Each trial began with a looming checkerboard circle in the center of the monitor. When participants had fixated on the checkerboard for one second,<sup>1</sup> a spoken word was presented diotically over headphones (Beyer DT-150) and visual icons corresponding to the four response choices (clip-art pictures of a beer, a pier, a deer, and a tear) were presented on the monitor (see Figure 4). For each listener, each response choice appeared in the same quadrant of the monitor on every trial at the onset of the auditory stimulus. The experiment was delivered under the control of E-prime experiment software (Psychology Software Tools, Inc.). Participants were instructed to find the picture of the word they heard and click it with the computer mouse as quickly as possible. The mouse click triggered the next trial. The stimuli in Blocks 1 through 3 were presented without breaks or any other overt demarcation; block structure was implicit and unknown to participants. There were 600 exposure trials (2 types (*beer/pier*, *deer/tier*) × 3 blocks × 10 exposure stimuli × 10 presentations) and 180 test trials (2 types (*beer/pier*, *deer/tier*) × 3 blocks × 3 F0 levels (high, mid, low) × 10 presentations). The entire session was completed in approximately 75 minutes.

**Analysis.** Our analyses focused on the effect of F0 on test-stimulus word recognition (i.e., the effect that F0 exerts on voicing categorization when VOT is perceptually ambiguous). We collected responses to mid-F0 (260 Hz) test stimuli to include variability along the F0 dimension so that high and low F0 were not too salient and to ensure that F0 has a gradient effect. For all cases, responses to the mid-F0 test stimuli indeed were intermediate the high- and low-F0 stimuli. Thus, the analyses focus on the high-versus low-F0 test stimuli to examine whether changes in the short-term diagnosticity of F0 for voicing categories influences how listeners use F0 in word recognition.

Because of the differences in category boundary across place-of-articulation, the *beer/pier* and *deer/tear* data are analyzed separately in 3 (Block: 1, 2, 3) × 2 (F0: high vs. low) repeated-measures ANOVAs on the average percent *pier* ([p]) responses and *tier* ([t]) responses, respectively. When an effect of F0 was

<sup>1</sup> This part of the procedure was attributable to the use of eyetracking in the experiment. The eye gaze data are not presented here because the data simply replicated the mouse-click data.

observed, changes in the degree of the F0 effect across blocks was investigated in planned post hoc *t* tests by examining the difference in voiceless ([p], [t]) responses across high F0 and low F0 conditions. Furthermore, to investigate the initial time course of F0 effect, responses to the first five of 10 presentations of test trials are examined within each block.

## Results

**Word recognition of exposure stimuli (Unambiguous VOTs).** Listeners' responses to exposure stimuli with unambiguous VOT values were examined to verify that participants used the primary cue, VOT, in word recognition (voiced: -10 ms for [b] and 0 ms for [d]; voiceless: 30 ms for [p] and 40 ms for [t]). Note that these VOT values were the center values and the most frequent in the distribution of voiced and voiceless VOTs among exposure stimuli (see Figure 2). Mean percentages of expected (correct) responses collapsed for *beer* and *deer* (voiced) and *pier* and *tier* (voiceless) were high: voiced,  $M = 94.9$ ,  $SE = .85$ ; voiceless,  $M = 99.0$ ,  $SE = .20$ , indicating that listeners indeed used VOT appropriately for word recognition.

**Word recognition of test stimuli (Ambiguous VOTs).** Figure 5<sup>2</sup> presents the mean percent voiceless (*pier* and *tear*) responses for high and low F0 across Blocks 1 to 3 (i.e., Canonical, Neutral and Reversed F0/VOT correlations). As is apparent in the figure, the influence of F0 decreased across blocks.

Three (Block: 1, 2, 3)  $\times$  2 (F0: high vs. low) repeated-measures ANOVAs were run on the average percent *pier* ([p]) responses and *tier* ([t]) responses, respectively. For [p] responses, there was a significant main effect of F0,  $F(1, 13) = 28.628$ ,  $p = .0001$ , and a significant interaction between Block and F0,  $F(2, 26) = 8.254$ ,  $p = .0017$ . The effect of Block was not significant,  $F(2, 26) = .465$ ,  $p = .6331$ . The significant F0  $\times$  Block interaction indicates that F0's influence on word recognition was not consistent across the three blocks. The mean difference scores in percent voiceless response across high- versus low-F0 stimuli were 42.1%, 43.6%, and 12.9% in Blocks 1, 2, and 3, showing a steep decrease in Block 3 when the F0/VOT correlation reversed. Across blocks, these

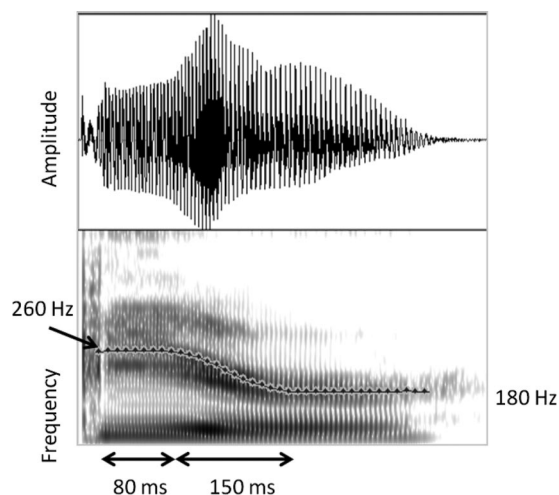


Figure 3. Waveform and spectrographic representation of a stimulus, pier, showing mid-F0 onset (260 Hz).



Figure 4. Images displayed on the computer monitor as response choices.

differences were compatible with long-term English experience with the F0/VOT correlation: the higher F0 stimuli elicited more voiceless responses to stimuli with ambiguous VOT values than did the lower F0 stimuli.

Paired-sample *t* tests indicated that F0 exerted a significant effect on word recognition for ambiguous-VOT test trials in Blocks 1 and 2, Block 1,  $t(13) = 5.462$ ,  $p = .0001$ ; Block 2,  $t(13) = 4.659$ ,  $p = .0004$ , but did not affect word recognition in Block 3,  $t(13) = 2.061$ ,  $p = .0599$ ; alpha adjusted to .017 for multiple comparisons. The significant interaction between the F0 effect and exposure blocks, as well as the results of post hoc tests, indicate that the extent to which listeners used F0 information in categorizing *beer* and *pier* changed over the course of the experiment. Specifically, the influence of F0 on voicing categorization diminished greatly in the final block in which listeners experienced a reversal in the canonical F0/VOT correlation.

The results for [t] closely matched those for [p]. There was a significant main effect of F0,  $F(1, 13) = 20.475$ ,  $p = .0006$ , a significant Block  $\times$  F0 interaction,  $F(2, 26) = 9.116$ ,  $p = .0010$ , and no significant main effect of Block,  $F(2, 26) = 1.100$ ,  $p = .3478$ . The mean difference scores in percent voiceless response as a function of F0 were 38.6%, 35.0%, and 16.4% in Blocks 1, 2, and 3, again showing a steep decrease in Block 3. Post hoc tests revealed a significant difference between the two F0 levels for the first two blocks, Block 1,  $t(13) = 6.232$ ,  $p = .0001$ ; Block 2,  $t(13) = 3.787$ ,  $p = .0023$ , and a marginally significant difference for Block 3,  $t(13) = 2.626$ ,  $p = .0209$ ;  $\alpha = .017$ . F0 continued to exert a significant influence on categorization in the final block, but the significant F0 by Block interaction in the omnibus

<sup>2</sup> Error bars are not plotted in graphs presenting data separately for within-subject conditions (i.e., percent voiceless for high and low F0 conditions). For within-subject designs, the consistency of the across-condition differences, and not the within-condition variability, contributes to statistical significance and so error bars on graphs can be misleading. Error bars are plotted in the graphs presenting difference scores computed from within-subject conditions (i.e., Figures 6, 9, and 11) because, in these cases, error bars indicating the variability of the difference scores are appropriate and meaningful.

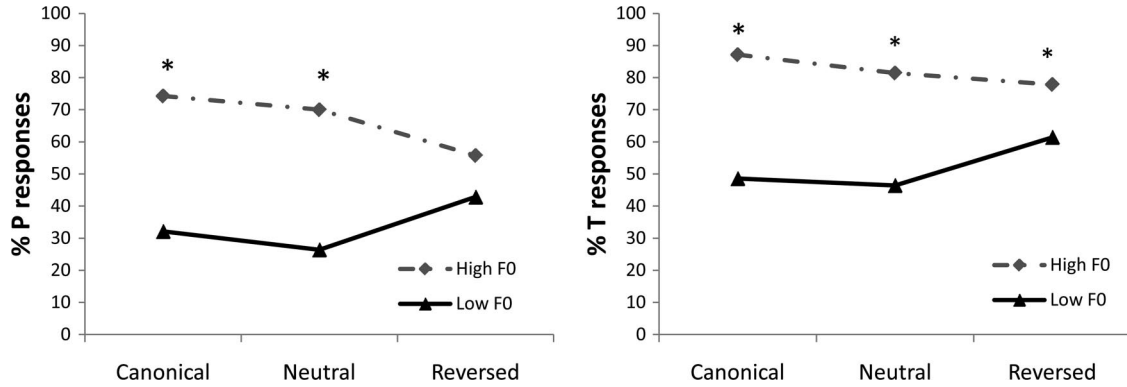


Figure 5. Percent voiceless responses for beer-pier series (left) and deer-tear series (right) across three exposure blocks (canonical, neutral, and reversed) in Experiment 1. Responses only to ambiguous test stimuli are plotted. Separate lines represent low-F0 (230 Hz) and high-F0 (290 Hz) conditions.

ANOVA demonstrates that the magnitude of F0's influence on categorization varied as a function of short-term experience with the dimension-based regularity.

To further investigate the changes in the magnitude of F0's influence for [t] responses, the degree of F0 effect, derived as a difference in voiceless responses between high F0 and low F0 test conditions, was examined across three blocks (see Figure 6). A repeated-measures ANOVA investigating the effect of Block returned a significant main effect of Block,  $F(2, 26) = 9.116, p = .001$ . Post hoc tests indicated that whereas the difference between Block 1 and Block 2 was not statistically significant, the difference was significant between Block 1 and Block 3,  $t(13) = 3.419, p = .005$ , and between Block 2 and Block 3,  $t(13) = 5.252, p < .000$  ( $\alpha = .017$ ). This analysis confirms that the influence of F0 was attenuated in the final block.

Finally, to investigate the time course of F0 effect at a finer scale, responses to the first five of 10 presentations of test trials (VOT-neutral *beer/pier* and *deer/tear* with high and low F0) were examined within each block. Recall that test trials were randomly interspersed among the exposure trials. On average, 10 exposure trials were presented before the first pairs of high-F0 and low-F0 test trials were presented (4.7 for [b/p] and 4.9 for [d/t]). Similarly,

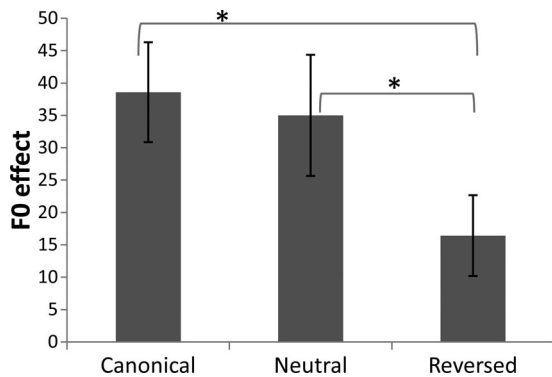


Figure 6. F0 effect (difference in percent voiceless responses between high and low F0 test trials) for deer-tear series across three exposure blocks (natural, neutral and reversed) in Experiment 1. Error bars indicate 1 standard error.

on average 30 exposure trials occurred within the span of presentation of the first two pairs of test trials; 50 exposure trials before the first three pairs of test trials; 70 exposure trials before the first four pairs of test trials; and 90 exposure trials before the first five pairs of test trials.

A series of  $3 \times 2$  (Block  $\times$  F0) repeated measures ANOVAs were conducted to compare voiceless responses (*pier* and *tear*) to the high- and low-F0 test stimuli in each exposure condition. The first ANOVA examined the responses to the first pair of high- and low-F0 test stimuli collapsed across [b/p] and [d/t] trials, the second test compared the responses to the first two pairs of the test stimuli, and so forth up to the first five pairs. The results of ANOVAs are reported in Table 1. A significant F0 effect and a marginally significant Block  $\times$  F0 interaction ( $p = .052$ ) were observed even for the first test trial pairs. Listeners' cumulative response after the second test trial pairs showed the eventual patterns observed by averaging across all 10 presentations (i.e., a significant F0 effect and significant Block  $\times$  F0 interaction), indicating that an influence of the perturbation of the F0/VOT relationship on word recognition emerged very quickly. On average, listeners had heard just 10 exposure trials ([b/p] and [d/t] combined) at this point.

### Discussion

These results demonstrate that listeners track relationships between acoustic dimensions in online speech processing, thus reshaping how information is evaluated in phonetic categorization and word recognition. Listeners are sensitive to short-term deviations in the correlation between acoustic dimensions (F0 and VOT) defining phonetic categories. In this experiment, experience with manipulations of the canonical English F0/VOT correlation caused listeners to down-weight their use of F0 in voicing categorization. Furthermore, exposure to the reversed F0/VOT correlation had an almost immediate influence on how listeners used F0 in word recognition. In the first two thirds of the experiment (Canonical and Neutral blocks), listeners more often categorized ambiguous VOT tokens with a higher F0 as voiceless (*pier* and *tear*). In the last third of the experiment, this correlation reversed among perceptually unambiguous tokens and listeners greatly reduced reliance on F0 as a cue to voicing categorization. The rate of listeners'

Table 1  
*F0 Effect in the First 5 Test Stimuli in Experiment 1*

Test pair(s)	Descriptive				ANOVA				
	Block	High F0	Low F0	Diff	Source	df 1	df 2	<i>F</i>	<i>p</i>
1	Canonical	1.57	0.79	0.79	Block	2	26	0.062	0.940
	Neutral	1.57	0.71	0.86	F0	1	15	17.333	0.001*
	Reversed	1.14	1.07	0.07	Block*F0	2	26	3.317	0.052
1 and 2	Canonical	3.21	1.79	1.43	Block	2	26	0.781	0.468
	Neutral	3.07	1.29	1.79	F0	1	15	22.114	0.000*
	Reversed	2.43	2.00	0.43	Block*F0	2	26	7.261	0.003*
1 to 3	Canonical	5.00	2.79	2.21	Block	2	26	1.149	0.333
	Neutral	4.57	2.21	2.36	F0	1	15	30.799	0.000*
	Reversed	3.93	2.93	1.00	Block*F0	2	26	5.268	0.012*
1 to 4	Canonical	6.57	3.79	2.79	Block	2	26	1.723	0.198
	Neutral	6.21	3.00	3.21	F0	1	15	38.158	0.000*
	Reversed	5.29	3.71	1.57	Block*F0	2	26	4.806	0.017*
1 to 5	Canonical	8.21	4.50	3.71	Block	2	26	0.966	0.394
	Neutral	7.79	3.86	3.93	F0	1	15	38.596	0.000*
	Reversed	6.50	4.93	1.57	Block*F0	2	26	7.211	0.003*

\* Significant at the  $p < .05$  level.

adaptation to the change in the relationship of the two acoustic dimensions was impressive: a detailed analysis of trials revealed that already at the *first* presentation of test words after the correlation reversal, there was a reweighting of F0 use. The mean number of reverse-correlation exposure stimuli experienced by listeners before the first test words was 10 (4.7 for [b/p] and 4.9 for [d/t]). Thus, with just five instances each of the reversed acoustic cue correlation, listeners had adjusted their use of the F0 dimension for speech categorization. This indicates a highly responsive and dynamic perceptual system adapting to the regularities of incoming speech. The down-weighting of F0 as a cue for speech categorization persisted through the remainder of the experiment.

Nonetheless, this rapid plasticity had limits. It is worth pointing out that although these data indicate that listeners track and make use of the regularity that resides *within* a speech segment, listeners were not completely susceptible to the statistical regularities of the immediate environment. If perception simply mirrored the short-term environment, we would expect the influence of F0 to mirror the correlation in the input; there would have been a full reversal in the influence of F0 on word recognition in Block 3 (a cross-over interaction). Whereas the reversed relationship of F0 to VOT in Block 3 greatly attenuated listeners' use of F0 in voicing judgments, it did not cause listeners to reverse their use of F0 relative to voicing. Instead, F0's influence on perception was down-weighted significantly. Thus, rapid dimension-based learning may be constrained by long-term regularities (such as the canonical English F0/VOT correlation) such that the local, short-term regularity does not override learned long-term patterns, even if the local pattern is systematic and robust.

Two aspects of the design of Experiment 1 bear note in interpreting this observation. First, the relationship of F0 to voicing categories varied across the experiment (Canonical to Neutral to Reversed), creating significant acoustic variability in F0 across the experiment. Previous research has shown that increasing variability along an acoustic dimension (F0 in this case) can affect listeners' perceptual weighting of cues along the dimension such that the influence of variable dimensions on categorization is attenuated

(Holt & Lotto, 2006). If, instead of tracking the correlation between F0 and voicing categories as it shifted, listeners down-weighted the influence of the variable F0 dimension, we may observe similar results. If information is perceptually weighted proportionate with its certainty, more variable sources of information may be relied upon less.

Second, the experiment included a Canonical condition (Block 1) with exposure to the typical English F0/VOT correlation to assess listeners' baseline use of F0 in voicing categorization. Given the evidence for sensitivity to dimension-based regularity observed in the results, one might question whether learning is evident even within Block 1. Exposure to the Canonical relationship between F0 and VOT may reinforce, and perhaps exaggerate, effects of the long-term English F0/VOT correlation. Said another way, although Experiment 1 sought evidence of learning in Block 3, it is possible that the perceptual patterns of Block 1 also indicate learning relative to an (unmeasured) baseline. If exposure to the natural correlation in Block 1 amplifies the influence of F0 on categorization responses, its influence may affect learning of the reversed correlation.

Thus, it is possible that a reversal in the influence of F0 on word recognition would appear when listeners experience this voice as producing only the reversed F0/VOT correlation. Furthermore, although limited exposure to the short-term correlation reversal in Experiment 1 was not sufficient to trigger a qualitative shift in the dimensions' relationship, it remains possible that greater experience would lead listeners to relate F0 to VOT in a manner consistent with short-term, but not long-term, experience.

## Experiment 2

Experiment 2 investigates this possibility by exposing listeners to only the reversed F0/VOT correlation in the course of the experiment. The numbers of exposure and test stimuli were equivalent with Experiment 1 (600 and 180, respectively), but all exposure stimuli sampled the reversed-correlation F0/VOT distributions.

## Method

**Participants.** Fifteen native-English listeners participated for a university credit or a small payment. No Experiment 1 listeners participated in this experiment, and all listeners were university students or employees. All reported normal hearing.

**Stimuli and procedure.** Stimuli for the Reversed correlation condition as well as the VOT-neutral test stimuli (see Figure 2) were used in this experiment. Listeners thus heard voiced stops with higher F0s and voiceless stops with lower F0s, opposite of the natural English pattern, throughout the experiment. Each of the exposure stimuli was presented 30 times in a random order. The VOT-neutral test stimuli were also presented 30 times each, interspersed randomly among the exposure stimuli. Trials proceeded continuously and listeners performed the word-recognition task throughout the experiment.

There were 600 exposure trials (2 types (*beer/pier*, *deer/tear*)  $\times$  10 exposure stimuli  $\times$  30 presentations) and 180 test trials (2 types (*beer/pier*, *deer/tear*)  $\times$  3 F0 levels (high, mid, low)  $\times$  30 presentations). The procedure was exactly the same as Experiment 1. The entire session was completed in approximately 75 minutes.

## Results

**Word recognition of Exposure Stimuli (Unambiguous VOTs).** Listeners' responses to the stimuli with unambiguous VOT values were examined to verify task compliance. Mean percentages of expected (correct) responses collapsed for [p/b] and [d/t] series were high (voiced:  $M = 93.6$ ,  $SE = 1.2$ ; voiceless:  $M = 98.1$ ,  $SE = .27$ ).

**Word recognition of Test Stimuli (Ambiguous VOTs).** Mean percent *pie*r ([p]) responses for high F0 and low F0 were 72.7% and 70.0% ( $SE = 5.6$  and  $6.2$ ), and mean percent *tear* ([t]) responses were 22.4% and 16.9% ( $SE = 7.5$  and  $7.0$ ). Because the F0/VOT correlation did not change across the experiment, there were no blocks in the sense of Experiments 1. Thus, the mean percent *pie*r and *tear* responses for high- and low-F0 levels collapsed for all 30 repetitions were compared. Separate paired  $t$  tests returned no significant difference due to F0 for either [p] or [t],  $t(14) = -.492$ ,  $p = .6301$ ;  $t(14) = -1.074$ ,  $p = .3008$ , indicating an overall lack of an effect of F0. Listeners ceased using F0 as a cue to categorizing VOT-ambiguous test stimuli.

To observe possible changes in listeners' responses over time, the data were divided into three subsets of equal numbers of exposure and test trials (10 repetitions of each in each subset), similar to the block structure of Experiment 1. Figure 7 shows the mean percent voiceless responses for two F0 levels (high and low) across these three phases of the experiment.

A 3 (Phase: 1, 2, 3)  $\times$  2 (F0: high vs. low) repeated-measures ANOVA run on percent [p] found no significant main effect for either of the factors, Phase,  $F(2, 28) = .987$ ,  $p = .3853$ ; F0,  $F(1, 14) = .242$ ,  $p = .6301$ ; and no interaction between the two,  $F(2, 28) = .691$ ,  $p = .5092$ . Thus, although there appears in Figure 7 to be some effect of F0 in the first third of the experiment, the percent voiceless difference attributable to F0 was not statistically significant, Phase 1,  $t(14) = -1.977$ ,  $p = .0681$ ; Phase 2,  $t(14) = .000$ ,  $p = 1.000$ ; Phase 3,  $t(14) = .000$ ,  $p = 1.000$  ( $\alpha = .017$ ). The results were similar for [t/d]. A 3 (Phase: 1, 2, 3)  $\times$  2 (F0: high vs. low) repeated-measures ANOVA run on percent [t] found no significant main effect for either factors, Phase,  $F(2, 28) = 1.277$ ,  $p = .2947$ ; F0,  $F(1, 14) = 1.154$ ,  $p = .3008$ , but it found a marginally significant interaction between F0 and Phase,  $F(2, 28) = 3.207$ ,  $p = .0557$ . Post hoc tests, however, indicated that these differences were not significant: Phase 1,  $t(14) = -2.467$ ,  $p = .0271$ ; Phase 2,  $t(14) = .295$ ,  $p = .7722$ ; Phase 3,  $t(14) = -.825$ ,  $p = .4231$  ( $\alpha = .017$ ). These results indicate no influence of F0 even in the first phase.

As for Experiment 1, response to the first five pairs of the 30 test stimuli were analyzed to examine possible early changes in the sensitivity to F0. A series of paired  $t$  tests was used to compare voiceless responses (*pie*r and *tear*) to the first pair (low F0 and high F0) of test trials and cumulatively up to the first five pairs (see Table 2). Although the F0 effect was absent in the first test pairs, it appeared in the subsequent response patterns. As the overall analysis indicated above, there was a possible trend of F0 effect in the first third of the experiment, but the effect was absent in the last two thirds as well as when responses were averaged across the course of experiment. It appears, therefore, that the F0 effect lingered even after listeners experienced about 90 exposure trials (the average number of exposure trials presented before the presentation of fifth test trial pairs). Note, however, that it is quite possible that the F0 effect was attenuated compared with what

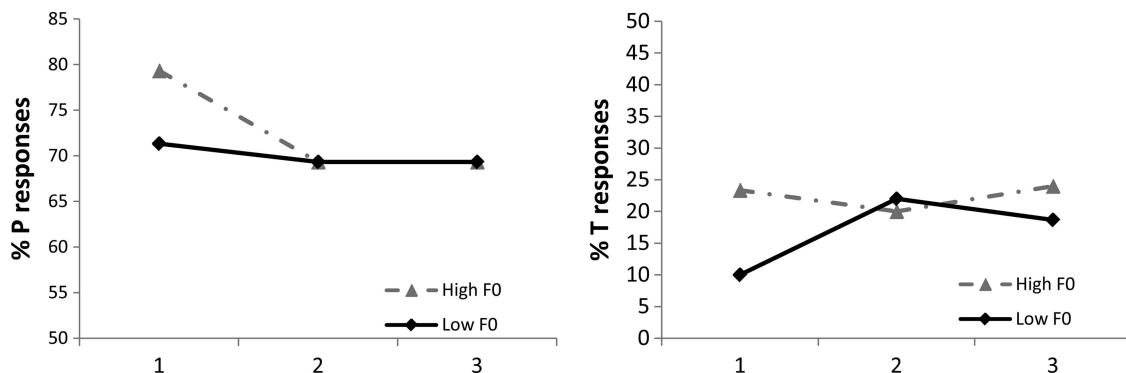


Figure 7. Percent voiceless responses for beer-pier series (left) and deer-tear series (right) across three phases of Experiment 2. Responses only to ambiguous test stimuli are plotted. Separate lines represent low-F0 (230 Hz) and high-F0 (290 Hz) conditions.

Table 2  
*F0 Effect in the First 5 Test Stimuli in Experiment 2*

Test pair(s)	Block	High F0	Low F0	Difference	<i>df</i>	<i>t</i>	<i>P</i>
1	Reversed	0.87	0.87	0.00	14	0.000	1.000
1 and 2	Reversed	2.20	1.60	0.60	14	2.806	0.014*
1 to 3	Reversed	3.40	2.60	0.80	14	3.055	0.009*
1 to 4	Reversed	4.33	3.33	1.00	14	2.562	0.023*
1 to 5	Reversed	5.33	4.07	1.27	14	3.106	0.008*

\* Significant at the  $p < .05$  level.

might have been before the exposure, but that the attenuation was not enough to eliminate the effect entirely.

## Discussion

In Experiment 2, listeners were exposed to the F0/VOT correlation that was opposite of the natural English pattern across 600 exposure trials for three times as many reversed exposure trials as in Experiment 1. In addition, there was no Canonical condition in this experiment that may have reinforced the long-term English F0/VOT correlation in Experiment 1. Despite the substantially greater amount of exposure and absence of reinforcement of the long-term correlation, response patterns were similar: listeners attenuated the weight for F0 as a cue to voicing relative to the effect of F0 on voicing categorization observed in many previous studies (e.g., Lisker, 1986; Francis et al., 2008), but their use of F0 in signaling voicing categories for word recognition did not reverse. In other words, although listeners heard about one hour of speech from a single talker with a very consistent “accent” in her use of F0/VOT, they did not fully adapt speech processing to mirror this systematic distributional pattern in the input.

## Experiment 3

Experiment 3 adopts a more rigorous test to investigate listeners’ persistent resistance to adapt to the reversed F0/VOT correlation. In this study, listeners experienced the Reversed F0/VOT correlation across five consecutive days. On the first day of testing, listeners completed a short categorization (Baseline pretest) to examine the initial F0 effect before exposure. Immediately after the Baseline pretest, listeners were exposed first to the Canonical correlation (200 exposure and 40 test trials) and then to the Reversed correlation (200 exposure and 40 test trials). The Canonical correlation was included to examine whether there was any learning in this condition relative to baseline. On the second through the fifth day of testing, listeners were exposed only to the Reversed correlation. Each day, they heard a number of trials equivalent with the Reversed condition of Experiment 2 (600 exposure and 120 test trials) for a total of 2600 exposure trials (200 on Day 1 and 2400 on Days 2–5) over more than 2 hours. In sum, listeners were tested with Baseline pretest, Canonical correlation, Reversed correlation (Day 1), and Reversed correlation (Days 2–4) across 5 days.

## Method

**Participants.** Thirteen native-English listeners participated for a university credit or a small payment. None of the listeners

participated in other experiments reported here within six months. All listeners were university students or employees. All reported normal hearing. One participant’s data were excluded from the analysis for failure to complete the sessions.

**Baseline pretest stimuli.** In a short perceptual test, listeners categorized stimuli varying perceptually from *beer* to *pier* in nine VOT steps (–20, –10, 0, 5, 10, 15, 20, 30, and 40 ms) and from *deer* to *tear* in nine VOT steps (–10, 0, 10, 15, 20, 25, 30, 40, and 50 ms). The VOT interval of 5 ms was used for the middle five steps around the boundary region (e.g., 0 ms to 20 ms for *beer-pier* series) to capture perceptual shift that occurs around the category boundary and 10 ms was used for the two steps of either end of the series. The stimuli also varied in the F0 onset frequency of the vowel in two levels (i.e., 230 Hz for low F0 and 290 Hz for high F0). Responses to these stimuli were used to examine the magnitude of the baseline effect of F0 on voicing categorization before experimental exposure to F0/VOT correlations.

**Exposure and test stimuli.** Exposure stimuli for the Canonical correlation and Reversed correlation condition (see Figure 2) were used for Day 1 of testing. Exposure stimuli for the Reversed correlation condition were used for Day 2 through Day 5 of testing. The VOT-neutral test stimuli with low and high F0 (i.e., low = 230 Hz, high = 290Hz) were used across 5 days of testing.

**Procedure.** On Day 1, listeners completed a baseline pretest as well as an exposure test with the Canonical and Reversed correlation stimuli. On Day 2 through Day 5, listeners completed an exposure test with the Reversed correlation stimuli.

**Baseline pretest.** Seated in front of a computer monitor in a sound-attenuated booth, listeners categorized 10 random presentations each of the baseline stimuli presented diotically over headphones (Beyer DT-150). The experiment was under the control of E-prime experiment software (Psychology Software Tools, Inc.). A total of 360 trials (9 VOT  $\times$  2 types (*beer/pier*, *deer/tear*)  $\times$  2 F0 levels  $\times$  10 repetitions) were presented, blocked for *beer/pier* and *deer/tear* types, and counterbalanced for the order of block presentation. Each trial presented the stimulus word through headphones as well as visual icons on the computer monitor corresponding to the two response choices (pictures of a beer and a pier or pictures of a deer and a tear). Each response choice appeared in the same location on the computer screen on every trial: Voiced choice on the left, voiceless choice on the right. Participants were instructed to respond quickly by pressing a computer key designated for response choice physically matching the relative location of the picture on the screen (the left key for voiced choice, the right key for the voiceless choice). Each response triggered the next trial. Each baseline pretest took no longer than 15 minutes.

**Exposure and test.** On Day 1, the exposure-and-test part of the experiment began immediately after the baseline pretest. Listeners were exposed to the Canonical (Block 1) and then the Reversed (Block 2) F0/VOT correlation with unambiguous VOT values (exposure stimuli), while being tested by perceptually ambiguous VOT stimuli (test stimuli) varying in F0. In each block, the exposure stimuli were presented 10 times each in random order. The ambiguous-VOT test stimuli were each presented 10 times per block, interspersed randomly among the exposure stimuli. Trials proceeded continuously, with a shift in the F0/VOT correlation half way, and listeners performed the same word-recognition task throughout the experiment. In this experiment, too, the block structure was not apparent in the nature of the task. Participants were not informed that the experiment was divided into separate blocks or that the nature of the acoustic cues would vary.

Each trial proceeded exactly as described for the baseline pretest, except that four response choices were shown on the screen (pictures of a beer, a pier, a deer and a tear) as in Experiment 1. The stimuli in Blocks 1 and 2 were presented without breaks or any other overt demarcation; block structure was implicit and unknown to participants. The test stimuli were not described to participants and they were not differentiated from exposure stimuli by task. There were 400 exposure trials (2 types [*beer/pier*, *deer/tear*]  $\times$  2 blocks  $\times$  10 exposure stimuli  $\times$  10 presentations) and 80 test trials (2 types [*beer/pier*, *deer/tear*]  $\times$  2 blocks  $\times$  2 F0 levels [high, low]  $\times$  10 presentations). The exposure-and-test session was completed in approximately 20 minutes.

The exposure-and-test session of Day 2 through Day 5 proceeded exactly as described for the exposure-and-test of Day 1, except that listeners were exposed only to the Reversed F0/VOT correlation. The exposure stimuli were presented 30 times each in random order. The VOT-neutral test stimuli were each presented 30 times per block, interspersed randomly among the exposure stimuli. Thus, there were 600 exposure trials (2 types [*beer/pier*, *deer/tear*]  $\times$  10 exposure stimuli  $\times$  30 presentations) and 120 test trials (2 types [*beer/pier*, *deer/tear*]  $\times$  2 F0 levels [high, low]  $\times$  30 presentations). The exposure-and-test session was completed in approximately 30 minutes.

## Results

**Baseline pretest.** A 9 (VOT: 9 steps)  $\times$  2 (F0: high vs. low) repeated-measures ANOVA was run separately on percent [p] and [t] response. The test for [p] returned significant main effects of both factors and a significant interaction between the factors: VOT,  $F(8, 88) = 191.912, p < .000$ ; F0,  $F(1, 11) = 39.890, p < .000$ ; VOT  $\times$  F0,  $F(8, 88) = 7.079, p < .000$ . Post hoc comparisons of high- and low-F0 at each VOT step indicated that F0 effect was significant at the 4th and 5th VOT steps (VOT = 5 ms, F0 effect = 36.9; VOT = 10ms, F0 effect = 24.6):  $t(11) = -4.284, p = .001$ ;  $t(11) = -3.645, p = .004$  (alpha level adjusted to .006 for multiple comparisons). The F0 effect was second largest at 10 ms of VOT (24.6), which was the VOT value used for the [p] test stimuli.

The test of [t] returned significant main effects of both VOT and F0, as well as a significant interaction between the two factors: VOT,  $F(8, 88) = 154.759, p < .000$ ; F0,  $F(1, 11) = 41.221, p < .000$ ; VOT  $\times$  F0,  $F(8, 88) = 8.786, p < .000$ . Post hoc comparisons

of high- and low-F0 at each VOT step indicated that F0 effect was significant at the 5th and 6th steps of VOT (VOT = 20 ms, F0 effect = 26.9; VOT = 25ms, F0 effect = 24.6):  $t(11) = -4.225, p = .001$ ;  $t(11) = -4.144, p = .002$  ( $\alpha = .006$ ).

These results indicate that listeners are indeed sensitive to F0 in categorizing voicing. The baseline F0 effects for the ambiguous-VOT test stimuli were found to be 24.6 (*beer/pier*) and 26.9 (*deer/tear*). It is noted that the VOT value used for *beer/pier* ambiguous-VOT test stimuli throughout this study (VOT = 10ms) did not elicit the largest F0 effect (observed at 5 ms), and the responses to this VOT value were biased toward voiceless.

**Word recognition of Exposure Stimuli (Unambiguous VOTs).** Listeners' responses to the stimuli with unambiguous VOT values were examined to verify task compliance. Mean percentages of expected (correct) responses collapsed for [p/b] and [d/t] series were high (voiced:  $M = 92.7, SE = .68$ ; voiceless:  $M = 96.4, SE = .49$ ).

**Word recognition of Test Stimuli (Ambiguous VOTs).** Figure 8 reports the mean percent voiceless responses to VOT-neutral stimuli for two F0 levels (low, high) across 5 days of testing. Baseline pretest included these VOT-neutral stimuli (10ms for [p] and 20 ms for [t]) for the same F0 values (230 and 290 Hz). Thus, the responses to these stimuli in the Baseline pretest are included in the figure as well as in the analysis.

A 7 (Block: Day 1 Baseline, Day 1 Canonical, Day1 Reversed, Days 2–5 Reversed)  $\times$  2 (F0: low vs. high) ANOVA run on percent [p] responses with repeated-measures on Block and F0 found significant main effects of Block,  $F(6, 66) = 2.878, p = .015$ , F0,  $F(1, 11) = 11.376, p = .006$ , and a significant interaction between Block and F0,  $F(6, 66) = 5.761, p < .000$ . The significant interaction indicates that the effect of F0 was not consistent across the blocks. Paired  $t$  tests revealed a significant F0 effect in the Baseline and Day 1 Canonical: Day 1 Baseline,  $t(11) = -3.645, p = .004$ ; Day 1 Canonical,  $t(11) = -4.393, p = .001$  ( $\alpha = .007$ ). However, the F0 effect was not significant in any of the Reversed blocks, indicating that listeners ceased using F0 as a cue to voicing with exposure to the reversed correlation.

Given the significant difference in percent [p] responses in the Baseline and Canonical blocks, the magnitude of the F0 effect (difference in [p] responses between high F0 and low F0) was compared across these two blocks. A  $t$  test returned nonsignificant results. Thus, although there was a trend of learning in Canonical Block (see Figure 8), this was not statistically verified.

A 7 (Block)  $\times$  2 (F0) ANOVA run on percent [t] responses with repeated-measures on Block and F0 found significant main effects of F0,  $F(1, 11) = 37.270, p < .000$ , and a significant interaction between Block and F0,  $F(6, 66) = 5.476, p < .000$ , and no effect of Block,  $F(1, 15) = 1.793, p = .114$ . Post hoc tests indicated that F0 effect was statistically significant in all blocks except for Day 1 Reversed Block: Day 1 Baseline,  $t(11) = -4.225, p = .001$ ; Day 1 Canonical,  $t(11) = -5.864, p < .000$ ; Day 2 Reversed,  $t(11) = -5.860, p < .000$ ; Day 3 Reversed,  $t(11) = -3.971, p = .002$ ; Day 4 Reversed,  $t(11) = -3.908, p = .002$ ; Day 5 Reversed,  $t(11) = -4.371, p = .001$  ( $\alpha = .007$ ). The difference attributable to F0, therefore, was significant in most of the Reverse conditions for [t]; however, the significant F0  $\times$  Block interaction in the omnibus ANOVA indicates that the magnitude of the effect of F0 on word recognition decreased with exposure to the reversed correlation in the Reversed blocks.

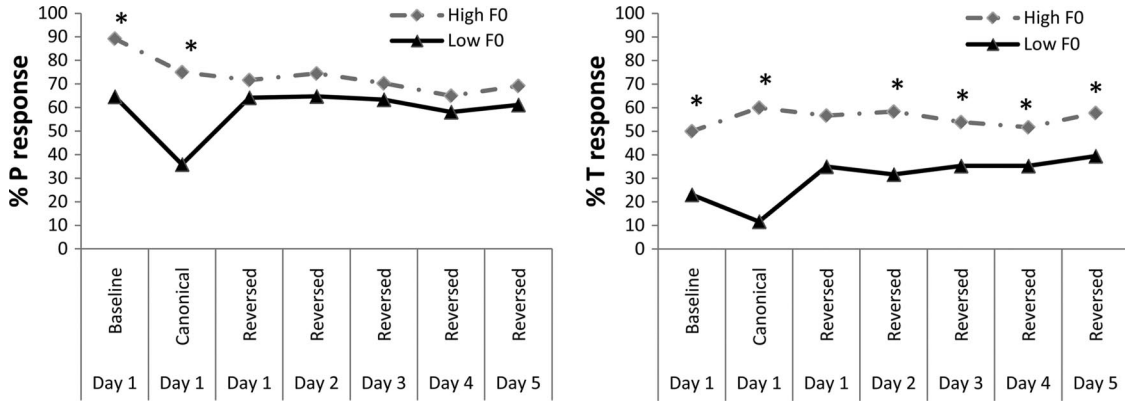


Figure 8. Percent voiceless responses for beer-pier series (left) and deer-tear series (right) across experimental blocks across 5 days in Experiment 3. Responses only to ambiguous test stimuli are plotted. Separate lines represent low-F0 (230 Hz) and high-F0 (290 Hz) conditions.

The change in the magnitude of the F0 effect in [t] responses was investigated by examining F0 effect measure (see Figure 9). First, a one-way ANOVA examined the mean F0-effect values of Reversed Blocks across 5 days and found no statistical differences. Given this, the data for the Reversed blocks were collapsed and compared with Canonical block and Baseline block. Post hoc tests found a statistically significant difference between Canonical block and Reversed blocks,  $t(11) = 3.994, p = .002$ ; no difference was found between Baseline and Canonical block or between Baseline and Reversed block ( $\alpha = .017$ ). Thus, for [t], we do observe attenuation of F0 effect (from Canonical to Reversed Blocks); however, F0 attenuation is weaker as reflected in a persisting F0 effect across the Reversed blocks (it was eliminated for [p]) and also as reflected in the results that there was no attenuation from Baseline to Reversed blocks.

In addition, responses to the first five of 10 presentations of test stimuli were examined within each block. On average, listeners experienced 10 exposure stimuli (4.7 for [b/p] and 4.9 for [d/t]) before the first pair of high-F0 and low-F0 test stimulus was presented. Similarly, 30 exposure stimuli occurred within the span of presenta-

tion of the first two pairs of test stimuli; 50 exposure stimuli before the first three pairs of test stimuli; 70 exposure stimuli before the first four pairs of test stimuli; and 90 exposure stimuli before the first five pairs of test stimuli.

A series of  $2 \times 2$  (Block  $\times$  F0) repeated measures ANOVAs was conducted to compare voiceless responses (*pier* and *tear*) to the high- and low-F0 test stimuli across Canonical Block (Day 1) and Reversed Block (Days 1–5). The first ANOVA examined the responses to the first pair of high- and low-F0 test stimuli collapsed across [b/p] and [d/t] trials, the second test compared the responses to the first two pairs of the test stimuli, and so forth up to the first five pairs across the Canonical Block (Day 1) and Reversed Block (Day 1). The results show that attenuation of F0 effect was present already in responses to the second pairs of test stimuli as reflected in a significant Block  $\times$  F0 interaction (see Table 3).

Additional  $2 \times 2$  (Block  $\times$  F0) ANOVAs were conducted to compare responses to high- and low-F0 test stimuli across Canonical correlation input (Day 1) and each of the Reversed correlation input on Day 2 through 5. Table 4 reports the results of first comparisons in which the Block  $\times$  F0 was significant. Whereas the attenuation of F0's influence on word recognition appeared quickly on Day 3 and Day 5 (it appeared in the first pairs of test stimuli), it appeared later on Day 2 and Day 4.

**Discussion**

Experiment 3 further verifies that listeners' use of F0 in signaling voicing categories for word recognition does not simply mirror experienced short-term regularities. Although listeners are highly sensitive to the reversal in the F0/VOT relationship, as evidenced by the down-weighting the influence of F0 on perception, experience with the "accented" reverse-correlation speech across about 4.5 hours over 5 days was not sufficient to remap voicing categories to reflect the experienced F0/VOT correlation. The long-term (Canonical) representation appears to be resilient to short-term perturbation that qualitatively changes dimensional relationships.

Across three experiments, listeners' short-term experience with "accented" speech possessing an F0/VOT relationship opposite

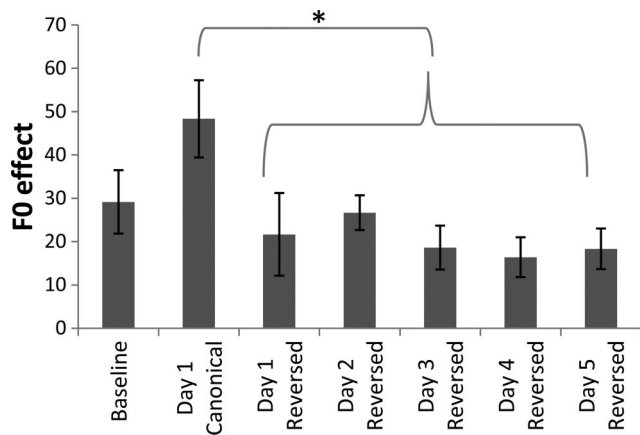


Figure 9. F0 effect (difference in percent voiceless responses between high and low F0 test trials) for deer-tear series across experimental blocks across 5 days in Experiment 3. Error bars indicate 1 standard error.

Table 3  
*F0 Effect in the First 5 Test Stimuli (Day 1 Canonical vs. Day 1 Reversed) in Experiment 3*

Test pair(s)	Descriptive				ANOVA				
	Block	High F0	Low F0	Diff	Source	df 1	df 2	<i>F</i>	<i>p</i>
1	Canonical	1.08	0.33	0.75	Block	1	11	3.062	0.108
	Reversed	1.17	1.00	0.17	F0	1	11	5.303	0.042*
					Block*F0	1	11	3.011	0.111
1 and 2	Canonical	2.33	0.92	1.41	Block	1	11	3.786	0.078
	Reversed	2.33	2.00	0.33	F0	1	11	10.569	0.008*
					Block*F0	1	11	7.406	0.020*
1 to 3	Canonical	3.50	1.33	2.17	Block	1	11	8.231	0.015*
	Reversed	3.67	3.08	0.59	F0	1	11	13.092	0.004*
					Block*F0	1	11	10.053	0.009*
1 to 4	Canonical	4.58	1.75	2.83	Block	1	11	10.028	0.009*
	Reversed	5.08	4.08	1.00	F0	1	11	16.484	0.002*
					Block*F0	1	11	14.955	0.003*
1 to 5	Canonical	6.08	2.08	4.00	Block	1	11	28.521	0.009*
	Reversed	6.17	5.08	1.09	F0	1	11	15.871	0.002*
					Block*F0	1	11	25.521	0.001*

\* Significant at the  $p < .05$  level.

that of long-term experience leads to rapid adjustment in how the F0 acoustic dimension contributes to word recognition. That this change occurs so quickly (within 5–10 trials of exposure) suggests perceptual processes that continuously track the distributional regularities across acoustic dimensions. However, it is unclear from the previous studies whether the observed decrease in F0's impact on word recognition arises because the dimension is perceptually down-weighted as a cue to voicing categorization in the word-recognition task, whether information along the dimension is simply no longer processed, or whether the F0/VOT relationship is remapped (albeit incompletely).

### Experiment 4

Experiment 4 attempts to tease apart these possibilities by exposing listeners to three experimental blocks, Canonical Correlation, Reversed Correlation, and Canonical Correlation. If listen-

ers continuously track the F0/VOT distribution and perceptually down-weight F0 in the Reversed Correlation block, we predict that the effect of F0 on ambiguous VOT test stimuli will rebound when the input regularity shifts back to Canonical correlation. If, however, listeners remap the relationship or cease to be sensitive to F0, more generally, upon experiencing Reversed Correlation, the attenuated F0 effect should persist into the second Canonical Correlation block.

### Method

**Participants.** Fifteen native-English listeners participated for a university credit or a small payment. None of the listeners participated in other experiments reported here within six months. All listeners were university students or employees. All reported normal hearing.

Table 4  
*F0 Effect in the First Test Stimuli (Day 1 Canonical vs. Day 2, 3, 4, and 5 Reversed) in Experiment 3*

Test pair(s)	Descriptive				ANOVA				
	Block	High F0	Low F0	Diff	Source	df 1	df 2	<i>F</i>	<i>p</i>
Day 2 1 to 5	Canonical	6.08	2.08	4.00	Block	1	11	9.122	0.012*
	Reversed	6.83	4.75	2.08	F0	1	11	33.708	0.000*
					Block*F0	1	11	7.036	0.022*
Day 3 1	Canonical	1.08	0.33	0.75	Block	1	11	3.477	0.089
	Reversed	1.00	1.00	0.00	F0	1	11	5.211	0.043*
					Block*F0	1	11	5.211	0.043*
Day 4 1 to 4	Canonical	4.58	1.75	2.83	Block	1	11	5.433	0.040*
	Reversed	4.58	3.58	1.00	F0	1	11	51.496	0.000*
					Block*F0	1	11	6.023	0.032*
Day 5 1	Canonical	1.08	0.33	0.75	Block	1	11	1.536	0.241
	Reversed	0.92	0.92	0.00	F0	1	11	5.211	0.043*
					Block*F0	1	11	11.88	0.005*

*Note.* The table shows the first test indicating a significant Block  $\times$  F0 interaction.

\* Significant at the  $p < .05$  level.

**Stimuli and procedure.** The baseline F0 effect before exposure was evaluated with pretest stimuli from Experiment 3. Exposure stimuli for the Canonical and Reversed condition as well as the VOT-neutral test stimuli with low and high F0 (i.e., low = 230 Hz, high = 290Hz) (see Figure 2) were used for the exposure test.

Seated in a sound attenuation booth, listeners first completed a baseline pretest, immediately followed by the exposure test. The procedure was the same as Day 1 in Experiment 3, except that in the exposure test, exposure stimuli began with Canonical (Block 1), shifted to Reversed (Block 2), and shifted back to Canonical (Block 3). In each block, the exposure stimuli were presented 10 times each in random order. The VOT-neutral test stimuli were each presented 10 times per block, interspersed randomly among the exposure stimuli. Trials proceeded continuously across three blocks and listeners performed the same word-recognition task throughout the experiment. Participants were not aware of the block structure or that the nature of acoustic cues would vary.

**Results**

**Baseline pretest.** A 9 (VOT: 9 steps) × 2 (F0: high vs. low) repeated-measures ANOVA was run separately on percent [p] and [t] response. The test for [p] returned significant main effects of both factors and a significant interaction between the factors: VOT,  $F(8, 112) = 290.431, p < .000$ ; F0,  $F(1, 14) = 77.517, p < .000$ ; VOT × F0,  $F(8, 112) = 107.097, p < .000$ . Post hoc comparisons of high- and low-F0 at each VOT step indicated that F0 effect was significant at the 4th and 5th VOT steps (VOT = 5 ms, F0 effect = 49.3; VOT = 10ms, F0 effect = 21.3):  $t(14) = -8.488, p < .000$ ;  $t(14) = -3.935, p = .002$  (alpha level adjusted to .006 for multiple comparisons). The F0 effect was second largest at 10 ms of VOT (21.3), which was the VOT value used for the [p] test stimuli.

The test of [t] returned significant main effects of both VOT and F0, as well as a significant interaction between the two factors: VOT,  $F(8, 112) = 241.019, p < .000$ ; F0,  $F(1, 14) = 48.696, p < .000$ ; VOT×F0,  $F(8, 112) = 13.717, p < .000$ . Post hoc comparisons of high- and low-F0 at each VOT step indicated that the F0 effect was significant at the fourth through seventh steps of VOT (VOT = 15 ms, F0 effect = 17.3; VOT = 20 ms, F0 effect = 25.3; VOT = 25 ms, F0 effect = 29.3; VOT = 30 ms, F0 effect = 8.0):

$t(14) = -5.773, p < .000$ ;  $t(14) = -4.672, p < .000$ ;  $t(14) = -4.725, p < .000$ ;  $t(14) = -3.292, p = .005$  (alpha level adjusted to .006 for multiple comparisons).

In this baseline pretest, too, listeners were sensitive to F0 in categorizing voicing. The baseline F0 effects for the VOT-neutral test stimuli were found to be 21.3 (*beer/pier*) and 25.3 (*deer/tear*), comparable to those found in Experiment 3 (24.6 and 26.9, respectively).

**Word recognition of Exposure Stimuli (Unambiguous VOTs).** Listeners' responses to the stimuli with unambiguous VOT values were examined to verify task compliance. Mean percentages of expected (correct) responses collapsed for [p/b] and [d/t] series were high (voiced:  $M = 96.1, SE = .24$ ; voiceless:  $M = 97.4, SE = .20$ ).

**Word recognition of Test Stimuli (Ambiguous VOTs).** Figure 10 reports the mean percent voiceless responses to VOT-neutral stimuli for two F0 levels (low, high) across three experimental blocks, Canonical 1, Reversed, and Canonical 2. Baseline pretest included these VOT-neutral stimuli (10ms for [p] and 20 ms for [t]) for the same F0 levels. Thus, the responses to these stimuli in the Baseline pretest are included in the figure as well as in the analysis.

Four (Block: 1, 2, 3, 4) × 2 (F0: high vs. low) repeated-measures ANOVAs were run on the average percent *pie* ([p]) responses and *tier* ([t]) responses, respectively. For [p] responses, there was a significant main effect of F0,  $F(1, 14) = 23.709, p < .000$ , and a marginally significant interaction between Block and F0,  $F(3, 42) = 2.739, p = .055$ . The effect of Block was not significant,  $F(3, 42) = 1.779, p = .166$ . The marginally significant F0 × Block interaction indicates that F0's influence on word recognition was not consistent across the four blocks. Post hoc tests indicated that the difference attributable to F0 was statistically significant in Baseline pretest, Canonical 1, and Canonical 2, but not in Reversed: Baseline,  $t(14) = -3.935, p = .001$ ; Canonical 1,  $t(14) = -3.697, p = .002$ ; Canonical 2,  $t(14) = -3.906, p = .002$  ( $\alpha = .013$ ). The results indicate that the F0 effect that was present in the Baseline test and in Canonical 1 exposure disappeared in Reversed block but rebounded in the final Canonical 2 block for *beer/pier* recognition.

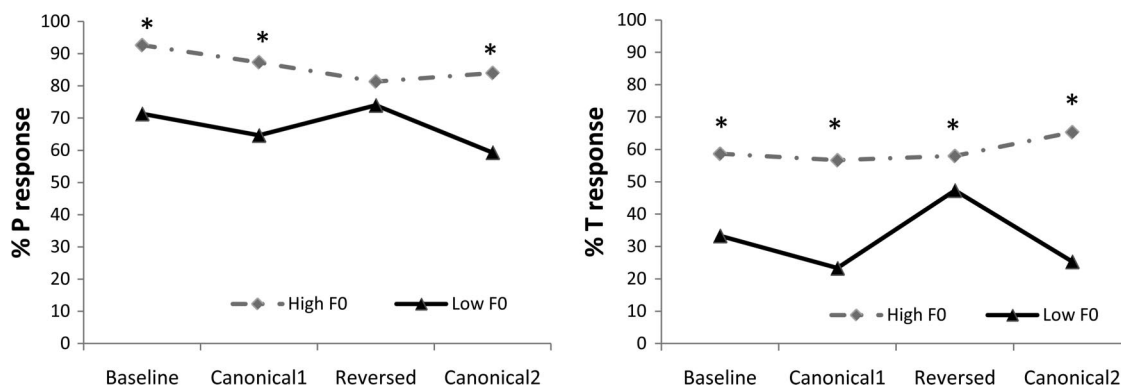


Figure 10. Percent voiceless responses for beer-pier series (left) and deer-tear series (right) across experimental blocks (baseline, canonical 1, reversed, canonical 2) in Experiment 4. Responses only to ambiguous test stimuli are plotted. Separate lines represent low-F0 (230 Hz) and high-F0 (290 Hz) conditions.

The results for [t] showed a significant main effect of F0,  $F(1, 14) = 44.593, p < .000$ , and a significant interaction between Block and F0,  $F(3, 42) = 7.010, p = .001$ . The effect of Block was not significant,  $F(3, 42) = 2.224, p = .099$ . Post hoc tests indicated that the difference attributable to F0 was statistically significant in all blocks: Baseline,  $t(14) = -4.672, p < .000$ ; Canonical 1,  $t(14) = -4.578, p < .000$ ; Reversed,  $t(14) = -3.552, p = .003$ ; Canonical 2,  $t(14) = -6.043, p < .000$  ( $\alpha = .013$ ).

Given the persistent F0 effect throughout the experiment, the magnitude of F0 effect in each block was compared (see Figure 11). A one-way ANOVA returned a significant effect of Block,  $F(3, 42) = 7.010, p = .001$ . Pairwise post hoc comparisons indicated that the difference between Canonical 1 and Reversed, as well as difference between Reversed and Canonical 2 were statistically significant: Canonical 1 versus Reversed,  $t(14) = 3.269, p = .006$ ; Reversed versus Canonical 2,  $t(14) = 5.537, p < .000$  ( $\alpha = .008$ ). These results indicate that although the F0 effect persists through the experiment, the magnitude of F0 effect was attenuated in the Reversed Correlation condition compared with two Canonical conditions.

Furthermore, responses to the first five of 10 presentations of test stimuli were compared between Canonical 1 and Reversed, as well as between Reversed and Canonical 2. On average, listeners experienced 10 exposure stimuli (4.7 for [b/p] and 4.9 for [d/t]) before the first pair of high-F0 and low-F0 test stimulus was presented. Similarly, 30 exposure stimuli occurred within the span of presentation of the first two pairs of test stimuli; 50 exposure stimuli before the first three pairs of test stimuli; 70 exposure stimuli before the first four pairs of test stimuli; and 90 exposure stimuli before the first five pairs of test stimuli.

A series of  $2 \times 2$  (Block  $\times$  F0) repeated measures ANOVAs were conducted to compare voiceless responses (*pier* and *tear*) to the high- and low-F0 test stimuli across Canonical 1 Block and Reversed Block. The first ANOVA examined the responses to the first pair of high- and low-F0 test stimuli collapsed across [b/p] and [d/t] trials, the second test compared the responses to the first two pairs of the test stimuli, and so forth up to the first five pairs. The results show that when the F0/VOT correlation shifted from Canonical to Reversed, attenuation of F0 effect appeared in responses to the third pairs of test stimuli as reflected in a significant Block  $\times$  F0 interaction (see Table 5).

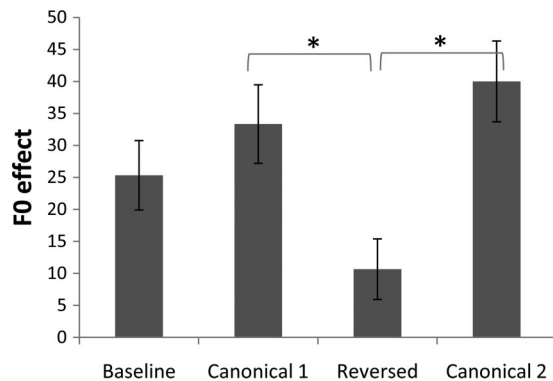


Figure 11. F0 effect (difference in percent voiceless responses between high and low F0 test trials) for deer-tear series across experimental blocks in Experiment 4. Error bars indicate 1 standard error.

Another series of  $2 \times 2$  repeated measure ANOVAs were conducted across Reversed Block and Canonical 2 Block. The results indicate that when the F0/VOT correlation shifted back to Canonical from Reversed, the increase of F0 effect appeared in responses to the fourth pair as reflected in the significant Block  $\times$  F0 interaction (see Table 6).

## Discussion

Listeners' reliance on the F0 in word recognition significantly decreased (or was eliminated in the case of *beer/pier*) when the input correlation shifted from Canonical to Reversed, but it increased again when the input correlation shifted from Reversed back to Canonical. Thus, listeners did not simply disregard the F0 dimension subsequent to exposure to the Reversed correlation. Instead, this experiment provides convincing evidence that perception continues to track changes in the F0/VOT correlation and adapt rapidly in response. These changes in response pattern occurred within an hour experimental session in response to implicit dimension-based distributional statistics in the speech input.

## General Discussion

There is mounting evidence of the adaptive plasticity of speech processing: listeners rely on local input regularities to dynamically "tune" long-term representations (e.g., Norris et al., 2003; Holt, 2005; Eisner & McQueen, 2005, 2006; Kraljic & Samuel, 2006, 2007; Clayards et al., 2008; Maye et al., 2008). In this way, 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 arising from talker-, accent-, and dialect-differences in spoken language. Understanding the nature of each of these facets of speech perception and their interaction is central to developing a full theoretical account.

An important implication of our findings is that reliance upon the very dimensions defining perceptual categories (e.g., F0, VOT) are dynamically, and rapidly, adjusted in online speech processing to accommodate regularities experienced in the ambient speech environment. The relationship of a particular acoustic dimension, or "feature," to phonetic categorization or word recognition is not rigidly fixed by long-term experience. Within just a few trials of exposure to a consistent deviation from the canonical native-language correlation between F0 and VOT dimensions, listeners in the present experiments had adaptively adjusted reliance on the F0 dimension in guiding word recognition. Thus these results indicate that the diagnosticity of a particular cue in speech categorization or word recognition is not simply a function of its particular value along an acoustic dimension but, rather, it is reliant on its relationship to other acoustic dimensions in short-term experience. Said another way, the feature space serving speech categorization and word recognition flexibly adjusts to local regularities.

We refer to this flexibility as *dimension-based statistical learning* to reflect that listeners have learned the relationship between coincident acoustic dimensions within a speech "object" (words in this case). This contrasts with investigations of statistical learning that primarily have focused, implicitly or explicitly, at the "object"

Table 5  
*F0 Effect in the First 5 Test Stimuli (Canonical 1 vs. Reversed) in Experiment 4*

Test pair(s)	Descriptive				ANOVA				
	Block	High F0	Low F0	Diff	Source	df 1	df 2	F	p
1	Canonical 1	1.47	.87	0.60	Block	1	14	.745	.403
	Reversed	1.47	1.20	0.27	F0	1	14	11.485	.004*
1 and 2	Canonical 1	3.13	1.80	1.33	Block*F0	1	14	2.059	.173
					Block	1	14	.189	.670
	Reversed	2.80	2.33	0.47	F0	1	14	21.000	.000*
1 to 3	Canonical 1	4.67	2.80	1.87	Block*F0	1	14	3.172	.097
					Block	1	14	.380	.547
	Reversed	4.07	3.73	0.33	F0	1	14	17.404	.001*
1 to 4	Canonical 1	6.13	3.80	2.33	Block*F0	1	14	7.747	.015*
					Block	1	14	.637	.438
	Reversed	5.47	5.00	0.47	F0	1	14	15.770	.001*
1 to 5	Canonical 1	7.47	4.67	2.80	Block*F0	1	14	7.487	.016*
					Block	1	14	1.672	.217
	Reversed	6.87	6.33	0.53	F0	1	14	15.625	.001*
					Block*F0	1	14	6.789	.021*

\* Significant at the  $p < .05$  level.

level whereby syllables, or phonetic categories, or words serve as the units across which regularities are computed (e.g., Saffran et al., 1996). The current results demonstrate that statistical learning across spoken language input is not limited to object-based regularities. Listeners also are sensitive to the relationship of physical dimensions that defines the objects and perception adapts flexibly to accommodate experienced changes in these relationships. An important goal for future research will be to examine how statistical learning of regularities between objects can be accomplished when the objects are defined by exemplars that are acoustically variable and probabilistically defined (see Emberson et al., 2009 and Lim et al., submitted for evidence that these learning challenges can be met simultaneously).

Dimension-based statistical learning may be particularly important for accommodating variability arising from nonnative-accented

speech because non-native speakers often use acoustic dimensions differently than do native speakers. Japanese speakers learning English vary second formant (F2) onset frequencies to distinguish English [r] and [l] although native English speakers primarily use the third formant (F3) to make this distinction (Lotto, Sato, & Diehl, 2004). Likewise, native-English learners of Korean rely on the English relationship between VOT and F0 in producing Korean stops, even though this relationship does not hold for Korean (Kim & Lotto, 2002). In such cases, adapting to the accented speech requires more than remapping within a stable perceptual space (Maye et al., 2008) or shifting a category boundary along an established dimension (Norris et al., 2003; Kraljic & Samuel, 2006, 2007). Rather, listeners must adjust how acoustic dimensions defining perceptual space relate to one another. This is just the sort of dimension-based statistical learning we observe in the present experiments.

Table 6  
*F0 Effect in the First 5 Test Stimuli (Reversed vs. Canonical 2) in Experiment 5*

Test pair(s)	Descriptive				ANOVA				
	Block	High F0	Low F0	Diff	Source	df 1	df 2	F	p
1	Reversed	1.47	1.20	0.27	Block	1	14	1.207	.290
	Canonical 2	1.47	0.87	0.60	F0	1	14	7.258	.017*
1 and 2	Reversed	2.80	2.33	0.47	Block*F0	1	14	1.522	.238
					Block	1	14	.516	.484
	Canonical 2	2.80	2.07	0.73	F0	1	14	6.248	.025*
1 to 3	Reversed	4.07	3.73	0.33	Block*F0	1	14	.365	.556
					Block	1	14	3.027	.104
	Canonical 2	4.07	2.93	1.13	F0	1	14	10.824	.005*
1 to 4	Reversed	5.47	5.00	0.47	Block*F0	1	14	2.565	.132
					Block	1	14	3.088	.101
	Canonical 2	5.67	3.80	1.87	F0	1	14	15.591	.001*
1 to 5	Reversed	6.87	6.33	0.53	Block*F0	1	14	6.274	.025*
					Block	1	14	5.845	.030
	Canonical 2	7.20	4.20	3.00	F0	1	14	26.825	.000*
					Block*F0	1	14	18.323	.001*

\* Significant at the  $p < .05$  level.

Whereas most studies demonstrating the adaptive plasticity of speech processing have examined the consequences of exposure to different speech patterns across groups' posttest perception (Norris et al., 2003; Eisner & McQueen, 2005; Kraljic & Samuel, 2007; Kraljic et al., 2008) or, more rarely, within listeners across pre- and postexposure tests (Kraljic & Samuel, 2006; Maye et al., 2008), the present work introduces a paradigm to evaluate learning surreptitiously as it unfolds in online speech processing. In these experiments, there was no overt difference between exposure and test trials or canonical and reverse correlation conditions. By embedding a small proportion of test trials among more-frequent exposure trials and maintaining a constant task, we were able to measure learning covertly and more continuously. In this way, it was possible to observe that listeners' sensitivity to the correlation change evolves rapidly, becoming evident within 10 trials of exposure to the new regularity in Experiment 1.

We argued in the introduction that dimension-based statistical learning may be particularly significant in speech processing because phonetic categories are inherently probabilistic and multidimensional. Nonetheless, dimension-based statistical learning is quite likely to be a general characteristic of perceptual processing within systems that must be sensitive to long-term regularities while remaining facile enough to accommodate deviations from these norms. Turk-Browne and colleagues (2008), for example, demonstrate that visual statistical learning is sensitive to both visual objects and to feature correlations defining objects in multidimensional visual sequences.

As final points, we wish to highlight several issues regarding the nature of the learning observed in the present experiments. A growing number of studies find evidence for a role for top-down feedback from the lexicon in influencing how listeners dynamically tune speech processing to accommodate distorted, accented, or otherwise altered speech input (e.g., Davis et al., 2005; Hervais-Adelman, Johnsrude, Davis, & Brent, 2002; Norris, McQueen, & Cutler, 2003; Kraljic & Samuel, 2006, 2007; Kraljic et al., 2008), and a debate has evolved around the nature of this learning (see McClelland, Mirman, & Holt, 2006; McQueen, Norris, & Cutler, 2006). In the present experiments, however, the response choices were all real English minimal pairs (i.e., *beer* vs. *pier* and *deer* vs. *tear*). Thus, lexical status did not provide information with which to resolve acoustically ambiguous input and could not serve as a teaching signal to drive learning.

Nonetheless, our exposure stimuli did possess consistent, perceptually-unambiguous VOT information with which to distinguish voiced and voiceless stops among exposure stimuli. For the majority of trials, F0 was superfluous to the task; the primary cue to voicing, VOT (Francis et al., 2008), reliably signaled consonant category membership (e.g., [b] or [p]) and therefore could guide highly accurate word recognition. Listeners' high word-recognition accuracy for exposure trials across experiments supports this supposition. However, despite the lack of an explicit task demand to use F0 in recognizing exposure trials, the results indicate that listeners did track its relationship to VOT across the exposure trials. Rather than a lexical "teacher," the reliable, unambiguous VOT information may have served as a signal to orient the relationship of the secondary, F0, dimension to voicing categories. The relationship of F0 to VOT was tuned through the reliable category information from VOT. Observing dimension-based statistical learning in the absence of lexical information

invites future study to investigate how lexical information would further shape dimension-based statistical learning and such studies may be helpful in resolving debates about the nature of perceptual learning in interactive versus feed-forward models (Massaro, 1998; McClelland & Elman, 1986; McClelland et al., 2006; McQueen et al., 2006; Norris, McQueen, & Cutler, 2000; Oden & Massaro, 1978) by providing evidence of interaction between the resolution of regularities in stimulus dimensions by the auditory system and various sources of information. Prior work has demonstrated that perceptual learning can be induced by visual information (Bertelson et al., 2003) or by phonotactic knowledge (Cutler et al., 2008). The current work demonstrates that the perceptual process can also use correlations between the very acoustic dimensions that define speech categories as a source of information to adjust perception. These results bid the question of 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.

As we noted above, the multiple acoustic dimensions that define speech categories are not equivalent in their perceptual weight; listeners rely on some much more than others (e.g., Idemaru & Guion-Anderson, 2010). In this regard, VOT is a stronger perceptual cue to English stop voicing than F0 (Francis et al., 2008). There is considerable overlap between F0 values for voiced and voiceless stops (see Figure 1), and F0 exerts the most influence in voicing categorization when VOT is ambiguous. This may have contributed to the effectiveness of the unambiguous VOT exposure stimuli in guiding learning in the present experiments. It will be important to investigate whether dimension-based learning is influenced by the strength of correlation between dimensions and the speech category, and by the perceptual weight of the dimension as a perceptual cue.

Consistently across experiments, the magnitude of learning was greater for *beer/pier* than for *deer/tear*. In the recognition of *beer* or *pier*, the influence of F0 was eliminated altogether in response to the Reversed correlation, whereas in the recognition of *deer* or *tier*, the F0 effect persisted at a much-diminished level. The cause for the difference between the *beer/pier* and *deer/tear* tasks is not clear at present. The baseline F0 effects were similar for both tasks as indicated by Experiment 3 and 4 (Experiment 3: *beer/pier* 24.6%, *deer/tear* 26.9%; Experiment 4: *beer/pier* 21.3%, *deer/tear* 25.3% at the Test VOT value). It is possible that the perceptually ambiguous VOT values chosen for the test stimuli had an influence in the magnitude of the learning observed: there was voiced bias for the VOT value (20 ms) of the *deer/tear* test stimuli whereas there was voiceless bias for the VOT value (10 ms) of the *beer/pier* test stimuli.

Another important point concerns the dual nature of speech processing noted above. Listeners in these experiments were not "blank slates" upon whom the statistical regularities of the experiment were written. Rather, the results indicate the importance of the interaction between sensitivity to long-term regularities in the native language and rapid adaptive plasticity to short-term online experience with perturbations away from these long-term regularities. In Experiment 1, listeners' use of F0 in word recognition changed significantly. Although exposure to the "accented" speech was short-term, the overall duration (approximately 25 minutes) is on par with or exceeds the duration of experience that elicits other kinds of statistical learning and adaptive plasticity for speech

perception among adults. Even 75 minutes of exposure in Experiment 2 and 30 minutes of exposure each day for 5 days in Experiment 3 did not result in behavior mirroring short-term input statistics (a reversal in the F0/VOT relationship). Instead, dimension-based statistical learning was evidenced by a strong down-weighting or elimination of the use of F0 in word recognition. The strong change in listeners' use of F0 in word recognition is evidence that they track dimensional relationships in online speech processing. However, it is notable that rather than mirroring the statistics of the immediate input, behavior exhibited a lingering influence of the long-term, dimension-based regularities of English.

In the present observations, a relatively more reliable perceptual source of information (unambiguous VOT) may adjust perception of a less-reliable source (F0) and perceptual decisions may have been made using all available information, including prior knowledge. Each of these points resonates with models of statistically optimal learning (see Fiser et al., 2010 for review). In using simultaneous information from multiple dimensions, information should be weighted commensurate with its certainty; less certain sources of information should be relied upon less. In the present experiments different F0 values were paired with perceptually unambiguous VOT values. Thus, F0 varied considerably within an experiment (or, in the case of Experiment 2 and 3, relative to long-term experience), making it a less certain source of information in distinguishing voicing categories. Consistent with a Bayesian framework, the diminishing effect of F0 on word recognition may have arisen from an interaction of long-term knowledge and a change in the informational content provided by the F0 dimension as a result of increased variability. Ultimately, models of dimension-based statistical learning will need to account for interactions between short-term and long-term regularities. For this, nonspeech category learning experiments for which all aspects of experience with category exemplars can be controlled and manipulated may be particularly helpful.

In sum, the diagnosticity of an acoustic dimension for perceptual categorization is relative to its relationship to the evolving distribution of regularity across time, not simply to its fixed value along the dimension. This perceptual tuning process is likely to be important for understanding how listeners deal with the acoustic perturbations in online speech input arising from accent, dialect, or dysarthria.

## 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* (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/>
- Castleman, W. A., & Diehl, R. L. (1996). Effects of fundamental frequency on medial and final [voice] judgments. *Journal of Phonetics, 24*, 383–398. doi:10.1006/jpho.1996.0021
- Chistovich, L. A. (1969). Variations of the fundamental voice pitch as a discriminatory cue for consonants. *Soviet Physics-Acoustics, 14*.
- 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
- Coleman, J. (2003). Discovering the acoustic correlates of phonological contrasts. *Journal of Phonetics, 31*, 351–372. doi:10.1016/j.wocn.2003.10.001
- 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.
- Davis, M. H., Johnsruide, I. S., Hervais-Adelman, A., Taylor, K., & McGettigan, C. (2005). Lexical information drives perceptual learning of distorted speech: Evidence from the comprehension of noise-vocoded sentences. *Journal of Experimental Psychology-General, 134*, 222–240. doi:10.1037/0096-3445.134.2.222
- 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
- Eisner, F., & McQueen, J. M. (2006). Perceptual learning in speech: Stability over time (L). *Journal of the Acoustical Society of America, 119*, 1950–1953. doi:10.1121/1.2178721
- Emberson, L. L., Liu, R., & Zevin, J. D. (2009). Statistics all the way down: How is statistical learning accomplished using novel, complex sound categories? In N. Taatgen and H. van Rijn (Eds.), *Proceedings of the 31st Annual Meeting of the Cognitive Science Society* (995–1000). Austin, TX: Cognitive Science Society.
- Fiser, J., Berkes, P., Orbán, G., & Lengyel, M. (2010). Statistically optimal perception and learning: From behavior to neural representations. *Trends in Cognitive Science, 14*, 119–130. doi:10.1016/j.tics.2010.01.003
- Francis, A. L., Baldwin, K., & Nusbaum, H. C. (2000). Effects of training on attention to acoustic cues. *Perception and 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. *The 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. *The Journal of the Acoustical Society of America, 47*, 613–617. doi:10.1121/1.1911936
- Haggard, M. P., Summerfield, Q., & Roberts, M. (1981). Psychoacoustical and cultural determinants of phoneme boundaries: Evidence from trading F-0 cues in the voiced–voiceless distinction. *Journal of Phonetics, 9*(1), 49–62.
- Hervais-Adelman, A., Johnsruide, I. S., Davis, M. H., & Brent, L. (2002). Adaptation to noise-vocoded speech in normal listeners: Perceptual learning depends on lexical feedback. In *Poster presented at the BSA Short Papers Meeting on Experimental Studies of Hearing and Deafness*, University of Sheffield, Sept 16th–17th.
- Hillenbrand, J. M., Clark, M. J., & Houde, R. A. (2000). Some effects of duration on vowel recognition. *The Journal of the Acoustical Society of America, 108*, 3013–3022. doi:10.1121/1.1323463
- Holt, L., & Wade, T. (2004). Non-linguistic sentence-length precursors affect speech perception: Implications for speaker and rate normalization. *Proceedings of From Sound to Sense: Fifty+ Years of Discoveries in Speech Communication*. Retrieved from <http://www.rle.mit.edu/soundtosense/conference/pdfs/fulltext/Friday%20Posters/FA-Holt-STS.pdf>
- Holt, L. L. (2005). Temporally non-adjacent non-linguistic sounds affect speech categorization. *Psychological Science, 16*, 305–312. doi:10.1111/j.0956-7976.2005.01532.x
- Holt, L. L., & Lotto, A. J. (2006). Cue weighting in auditory categorization: Implications for first and second language acquisition. *The Journal*

- of the *Acoustical Society of America*, 119, 3059–3071. doi:10.1121/1.2188377
- Idemaru, K., & Guion-Anderson, S. (2010). Relational timing in the production and perception of Japanese singleton and geminate stops. *Phonetica*, 67, 25–46. doi:10.1159/000319377
- Iverson, P., & Kuhl, P. K. (1995). Mapping the perceptual magnet effect for speech using signal detection theory and multidimensional scaling. *The Journal of the Acoustical Society of America*, 97, 553–562. doi:10.1121/1.412280
- 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
- Jongman, A., Wayland, R., & Wong, S. (2000). Acoustic characteristics of English fricatives. *The Journal of the Acoustical Society of America*, 108, 1252–1263. doi:10.1121/1.1288413
- Kim, M. R., & Lotto, A. J. (2002). An investigation of acoustic characteristics of Korean stops produced by non-heritage learners. *The Korean Language in America*, 7, 177–188.
- Kingston, J., & Diehl, R. L. (1994). Phonetic knowledge. *Language*, 70, 419–454. doi:10.2307/416481
- Kluender, K. R., & Walsh, M. A. (1992). Amplitude rise time and the perception of the voiceless affricate/fricative distinction. *Perception & Psychophysics*, 51, 328–333. doi:10.3758/BF03211626
- Kohler, K. J. (1982). F0 in the production of lenis and fortis plosives. *Phonetica*, 39, 199–218. doi:10.1159/000261663
- Kohler, K. J. (1984). Phonetic explanation in phonology: The feature fortis/lenis. *Phonetica*, 41, 150–174. doi:10.1159/000261721
- Kohler, K. J. (1985). The perception of lenis and fortis plosives in French. A critical re-evaluation. *Phonetica*, 42, 116–123. doi:10.1159/000261742
- Kraljic, T., Brennan, S. E., & Samuel, A. G. (2008). Accommodating variation: Dialects, idiolects, and speech processing. *Cognition*, 107, 51–81. doi:10.1016/j.cognition.2007.07.013
- Kraljic, T., & Samuel, A. G. (2006). Generalization in perceptual learning for speech. *Psychonomic Bulletin and 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
- Lim, S., Lacerda, F., & Holt, L. L. (submitted). Discovering functional units in continuous speech.
- Lisker, L. (1986). "Voicing" in English: A catalogue of acoustic features signaling /b/ versus /p/ in trochees. *Language and speech*, 29, 3–11.
- 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/. *Proceedings of From Sound to Sense: Fifty+ Years of Discoveries in Speech Communication*. Retrieved from <http://www.rle.mit.edu/soundtosense/conference/pdfs/fulltext/Saturday%20Posters/SB-Lotto-STS.pdf>
- Massaro, D. W. (1998). *Perceiving talking faces: From speech perception to a behavioral principle*. The MIT Press.
- Maye, J., Aslin, R. N., & Tanenhaus, M. K. (2008). The Weckud Wetch of the West: Lexical adaptation to a novel accent. *Cognitive Science: A Multidisciplinary Journal*, 32, 543–562.
- 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
- Mayo, C., & Turk, A. (2005). The influence of spectral distinctiveness on acoustic cue weighting in children's and adults' speech perception. *The Journal of the Acoustical Society of America*, 118, 1730–1741. doi:10.1121/1.1979451
- McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, 18, 1–86. doi:10.1016/0010-0285(86)90015-0
- 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
- McQueen, J. M., Norris, D., & Cutler, A. (2006). Are there really interactive processes in speech perception? *Trends in Cognitive Sciences*, 10, 533. doi:10.1016/j.tics.2006.10.004
- Narayan, C. R. (to appear). Developmental perspectives on phonological typology and sound change. In A. C. L. Yu (Ed.), *Origins of sound patterns: Approaches to phonologization*. Oxford, UK: Oxford University Press.
- Newport, E. L., & Aslin, R. N. (2004). Learning at a distance I. Statistical learning of non-adjacent dependencies. *Cognitive Psychology*, 48, 127–162.
- Nittrouer, S. (2004). The role of temporal and dynamic signal components in the perception of syllable-final stop voicing by children and adults. *The Journal of the Acoustical Society of America*, 115, 1777–1790. doi:10.1121/1.1651192
- Norris, D., McQueen, J. M., & Cutler, A. (2000). Merging information in speech recognition: Feedback is never necessary. *Behavioral and Brain Sciences*, 23, 299–325. doi:10.1017/S0140525X00003241
- 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
- Oden, G. C., & Massaro, D. W. (1978). Integration of featural information in speech perception. *Psychological Review*, 85, 172–191. doi:10.1037/0033-295X.85.3.172
- Polka, L., & Strange, W. (1985). Perceptual equivalence of acoustic cues that differentiate /r/ and /l/. *The Journal of the Acoustical Society of America*, 78, 1187–1197. doi:10.1121/1.392887
- 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
- Turk-Browne, N. B., Isola, P. J., Scholl, B. J., & Treat, T. A. (2008). Multidimensional visual statistical learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34, 399–407. doi:10.1037/0278-7393.34.2.399
- Whalen, D. H., Abramson, A. S., Lisker, L., & Mody, M. (1993). F0 gives voicing information even with unambiguous voice onset times. *The Journal of the Acoustical Society of America*, 93, 2152–2159. doi:10.1121/1.406678

Received August 6, 2010

Revision received March 23, 2011

Accepted May 23, 2011 ■