American Psychological Association

Journal of Experimental Psychology:
Human Perception and Performance

0096-1523/18/$12.00

2018, Vol. 44, No. 11, 1760–1779
http://dx.doi.org/10.1037/xhp0000569

Simultaneous Tracking of Coevolving Distributional Regularities in Speech

Xujin Zhang and Lori L. Holt
Carnegie Mellon University

Speech processing depends upon mapping variable acoustic speech input in a manner that reflects the long-term regularities of the native language. Yet, these mappings are flexible such that introduction of short-term distributional regularities in speech input, like those arising from foreign accents or talker idiosyncrasies, leads to rapid adjustments in the effectiveness of acoustic dimensions in signaling phonetic categories. The present experiments investigate whether the system is able to track simultaneous short-term distributional statistics present in speech input or if, instead, the global regularity jointly defined by these distributions dominates. Three experiments establish that adult listeners are able to track distinct simultaneously evolving regularities across time, given information to support the "binning" of acoustic instances. Both voice quality and visual information to indicate talker supported tracking of coevolving distributional regularities, even when the regularities are opposing and even when the acoustic speech tokens contributing to the distinct distributions are identical. This indicates that reweighting of perceptual dimensions in response to short-term regularities in speech input is not simply an accumulation of acoustic instances. Rather, the system is able to track multiple context-sensitive regularities simultaneously, with rapid context-dependent adaptive adjustments in how acoustic speech input maps to phonetic categories.

Public Significance Statement
We often encounter talkers whose speech departs from the norm due to a foreign accent or speech idiosyncrasy. Research demonstrates that listeners rapidly adapt to unusual or unexpected patterns of speech. But what is involved in this adaptation? The present experiments reveal that the perceptual system quickly adjusts the effectiveness of specific features of the speech input (like voice pitch) in signaling speech sounds like /b/ or /p/ that differentiate *beer* from *pier*. These adjustments are initiated when the detailed correlations among acoustic properties of speech shift in the input. Surprisingly, the system is not fooled when speech input blends competing correlations. Instead, listeners use information about voice and talker to simultaneously track coevolving correlations in speech input.

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

All perceptual systems face the competing demands of maintaining relatively stable representations that reflect long-term regularities experienced in the environment and adapting flexibly to short-term distributional regularities as they arise. This tension is very evident in perception of speech because the mapping of acoustic speech input to phonetic categories is complicated by the highly variable nature of speech across speakers, foreign accents, and regional dialects. For example, an English talker from Long Island, New York, may pronounce /s/ in *street* more similarly to /ʃ/ as in *sheet* (Kraljic, Brennan, & Samuel, 2008). Or, a native Italian speaker may produce English /I/, as in *chick*, with a vowel more similar to English /i/, as in *cheek* (Liu & Holt, 2015). This creates circumstances in which a particular acoustic signal may be consistent with multiple phonetic categories, or a particular phonetic category may have distinct acoustic realizations.

Rapid adaptation, or recalibration, of speech perception is one factor that contributes to effective speech processing in the face of these complicated mappings. A growing literature documents that speech perception adapts to acoustic regularities experienced across short-term input. These studies demonstrate that exposure to perceptually ambiguous speech acoustics in the context of information that disambiguates its mapping leads to rapid, online adjustments such that later perception of the ambiguous speech is shifted even when the disambiguating information is no longer present. This disambiguating information may arise from lexical (Kraljic & Samuel, 2005, 2006; Norris, McQueen, & Cutler, 2003), visual (Bertelson, Vroomen, & De Gelder, 2003; Vroomen,

This article was published Online First October 1, 2018.

Xujin Zhang and Lori L. Holt, Department of Psychology, and the Center for the Neural Basis of Cognition, Carnegie Mellon University.

Xujin Zhang is now at Google, Inc., Mountain View, California.

Research was supported by the National Institutes of Health (R01DC004674). Thanks to Christi Gomez for support in testing human participants.

Correspondence concerning this article should be addressed to Lori L. Holt, Department of Psychology, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213. E-mail: loriholt@cmu.edu

van Linden, de Gelder, & Bertelson, 2007), or acoustic (Idemaru & Holt, 2011, 2014; Schertz, Cho, Lotto, & Warner, 2016) sources. For example, lexical knowledge can push an acoustically ambiguous English consonant between /b/ and /p/ to be more often categorized as /b/ in the context of *_eef*, but as /p/ in the context of *_eace* (Ganong, 1980). Or, a silent video of a speaker articulating *aba* can shift categorization of an acoustically ambiguous consonant between /b/ and /d/ to /b/ whereas the same acoustic signal is heard as /d/ in the context of a face articulating *ada* (Bertelson et al., 2003). Repeated exposure to ambiguous acoustics that are resolved in this way results in shifts in listeners' speech categorization that persist even when the lexical or visual supporting information is no longer present. Perceptual adjustments like these have been referred to as *phonetic recalibration*, *phonetic retuning*, *perceptual learning*, or *adaptive plasticity* (Guediche, Blumstein, Fiez, & Holt, 2014; Norris et al., 2003; Samuel & Kraljic, 2009; Vroomen & Baart, 2009). These effects appear to be evident at early perceptual stages of processing (Trude & Brown-Schmidt, 2012) rather than arising from a response bias at the decision stage (Clarke-Davidson, Luce, & Sawusch, 2008).

Recently, a number of studies have demonstrated adaptive plasticity in how acoustic dimensions are mapped to phonetic categories (Idemaru & Holt, 2011, 2014; Lehet & Holt, 2017; Liu & Holt, 2015; Schertz et al., 2016), providing an especially useful means of understanding adaptive plasticity at these early stages of processing. Speech categories are defined by multiple, probabilistic acoustic dimensions, and some dimensions are more strongly associated with phonetic category membership than others. For example, in English, voiced stop consonants, such as /b/, are usually produced with a short voice onset time (VOT) and a lower fundamental frequency (F0), whereas voiceless stop consonants, such as /p/, are usually produced with a longer VOT and a higher F0 (Kingston & Diehl, 1994). Listeners use both VOT and F0 to categorize /b/–/p/. Yet, these acoustic dimensions differ in their effectiveness in signaling /b/ and /p/ categories. Perhaps because the distribution of natural speech productions overlaps considerably more along the F0 dimension than the VOT dimension in long-term experience (Idemaru & Holt, 2011), listeners rely more on VOT than F0 in perceptual categorization. They give VOT more *perceptual weight*, as evidenced by a stronger correlation between VOT and categorization responses when VOT and F0 covary across a two-dimensional acoustic space (Abramson & Lisker, 1985; Francis, Kaganovich, & Driscoll-Huber, 2008; Holt & Lotto, 2006; Idemaru & Holt, 2011). However, when VOT is perceptually ambiguous, listeners rely on the secondary dimension, F0, such that lower F0s signal /b/, whereas higher F0s signal /p/ (Castleman & Diehl, 1996; Whalen, Abramson, Lisker, & Mody, 1993). Although VOT is the dominant, more heavily perceptually weighted, dimension in /b/–/p/ categorization F0 also contributes as a secondary dimension. Further, the influence of F0 on /b/–/p/ categorization reflects the relationship between F0 and VOT in English speech production (i.e., higher F0 leads to more /p/ categorizations, whereas lower F0 results in more /b/ categorizations). In this way, adult speech categorization reflects the long-term distributional regularities of speech input.

Yet, speech input does not always mirror these long-term regularities in the short term. Listeners encounter talkers with dialects, accents, or speech idiosyncrasies that create short-term acoustic dimension regularities that depart from long-term experience. For example, due to the structure of Korean, a Korean talker learning English as a second language may produce English /b/–/p/ with an F0 × VOT correlation misaligned with the regularity typical of English (Kim & Lotto, 2002).

Idemaru and Holt (2011, 2014) first investigated how the perceptual system adapts in response to short-term deviations in acoustic dimensional regularities. In these studies, native-English listeners categorized /b/ versus /p/ in minimal pairs of words (rhymes *beer/pier*, *deer/tear*). Key to the design, the majority of stimuli were either sampled from an acoustic F0 × VOT space in a manner that aligned with the canonical English F0 × VOT correlation or violated it with a correlation reversal. In the latter case, this "artificial accent" paired longer VOTs with *lower* F0s and shorter VOTs with *higher* F0s contra the typical English regularity. Across both the canonical and the reverse sampling, VOT was always available to unambiguously signal /b/–/p/ category membership; only its relationship with F0 shifted upon introduction of the artificial accent. In this way, there was no strong pressure for listeners to rely on F0 for speech categorization. Even though F0 was not strictly necessary for speech categorization, Idemaru and Holt found that exposure to the artificial accent resulted in rapid down-weighting in listeners' reliance on F0 in /b/–/p/ categorization. Likewise, reintroduction of the canonical English F0 × VOT correlation resulted in rapid reestablishment of F0 as effective in signaling /b/ versus /p/.

Like other demonstrations of adaptive plasticity in speech perception, this *dimension-based statistical learning* demonstrates that short-term input regularities impact how speech input maps to long-term representations. Here, the heavily perceptually weighted VOT dimension disambiguates incoming speech input, allowing for unambiguous /b/–/p/ categorization and leading to online adjustments such that subsequent perception of ambiguous speech is shifted even when the disambiguating VOT information was no longer present among test stimuli. Dimension-based statistical learning also has been observed for vowel categorization across spectral and temporal acoustic dimensions (Liu & Holt, 2015) and occurs across both online word (Idemaru & Holt, 2011, 2014; Lehet & Holt, 2017; Liu & Holt, 2015) and nonword (Liu & Holt, 2015) recognition.

Given that it is a perceptually unambiguous *acoustic* dimension that drives these effects, it is tempting to view the learning as arising from a simple, low-level accumulation of distributional experience across acoustic input dimensions. Yet, when adaptive plasticity is driven by lexical information that disambiguates speech input, adjustments are organized by their episodic properties and are contextually sensitive (Samuel & Kraljic, 2009 for review). For example, if there is information to indicate that the ambiguity in speech input is characteristic of the talker producing it, then adaptive shifts in speech categorization are observed. If, instead, the speech ambiguity is an incidental consequence of some other factor then the system does not adapt. As an example, the presence of an external factor, like a pen in a speaker's mouth, blocks adaptive adjustments to speech categorization (Kraljic et al., 2008).

The present studies use dimension-based statistical learning to examine whether adaptive plasticity in speech perception is driven by an accumulation of global acoustic input regularities, or whether it is impacted by information that provides the opportunity to discover distinct local regularities within the global acoustic

statistics. In doing so, the studies also test whether the system is able to track, and adapt to, multiple regularities unfolding together. This is important in that speech input often involves multiple talkers. Although prior studies have addressed the extent to which adaptive plasticity is talker specific (e.g., Eisner & McQueen, 2005; Kraljic & Samuel, 2006), these studies have tended to examine generalization of learning across speech from one talker to speech from another talker, rather than the tracking of multiple simultaneously evolving short-term regularities (but see Kraljic & Samuel, 2007; Kraljic, Samuel, & Brennan, 2008; Trude & Brown-Schmidt, 2012).

An advantage of investigating the adaptive plasticity of speech perception using dimension-based statistical learning is that it provides a direct measure of how online recalibration to short-term input statistics impacts the effectiveness of specific acoustic dimensions in signaling speech categories. This contrasts with adaptive plasticity in speech categorization driven by lexical or visual information (e.g., Norris et al., 2003; Bertelson et al., 2003), which are evidenced as shifts in categorization boundaries without the availability of detailed information about the source of these shifts. In the present studies, we capitalize on the ability to precisely manipulate the local and global input regularities in speech input in the Idemaru and Holt (2011) dimension-based statistical learning paradigm. With it, we measure the direct impact of competing short-term speech input regularities on the effectiveness of acoustic dimensions in signaling speech categories. This allows us to advance current understanding through examination of several questions central to the representation of speech and its online tuning.

First, to what extent do the global distributional regularities present in short-term speech input influence the effectiveness of acoustic dimensions in signaling speech categories? Earlier, we noted that English speech tends more often to be perceived as /p/ than /b/ when spoken with a high F0. But, a male's high F0 may be equivalent in frequency to a female's low F0. What counts as "high" to the perceptual system? Experiment 1 examines whether the acoustic F0 dimension signals /b/–/p/ category membership in a manner *relative to the range of F0* experienced in a particular context.

Second, is it possible for the system to track simultaneous, coevolving short-term distributional statistics in the input, or is categorization impacted only by the global distributional statistics defined jointly by the local acoustic regularities? Third, if it is possible to track multiple regularities, what kind of information does the system need to effectively "bin" short-term statistics mixed in experience across time? Experiments 2 and 3 addressed these latter questions. In Experiment 2, two speech stimulus sets—each with a distinct voice—sample F0 and VOT values with opposing short-term F0 × VOT distributional statistics. In Experiment 3, a speech stimulus set with a common voice samples these same opposing regularities. But, the voice is paired with silent videos of two different individuals, each articulating in synchrony with one of the distinct samplings of distributions across the F0 × VOT acoustic space.

If dimension-based statistical learning is purely an accumulation of low-level acoustic dimensional regularities across time then the global short-term distributional regularities, which include opposing F0 × VOT correlations across the manipulations in voice (Experiment 2) and visual information (Experiment 3), will "cancel" one another providing no basis for dimension-based statistical learning. However, if voice quality or visual information is sufficient to support the "binning" of the distributions present in short-term speech input then we predict patterns consistent with the down-weighting of F0 apparent in dimension-based statistical learning upon introduction of an "artificial accent," even as the canonical English regularity is maintained across acoustic speech associated with the other voice or face. This would indicate that the system is able to track multiple regularities simultaneously, with the possibility of context-dependent adaptive adjustments in how acoustic speech input maps to phonetic categories. In this way, the present experiments use the dimension based statistical learning paradigm of Idemaru and Holt (2011) to determine the extent to which listeners track simultaneous coevolving distributional regularities in speech input.

Experiment 1

Prior research has demonstrated that the F0 of a following vowel affects /b/–/p/ categorization (e.g., Abramson & Lisker, 1985; Whalen et al., 1993). When the vowel has a higher F0, a stop consonant is more often categorized as /p/ than when the vowel has a low F0. Of course, a low F0 for a female talker may be a high F0 for a male talker. The purpose of Experiment 1 is to establish the extent to which dimension-based statistical learning is observed across distributions that sample different ranges of F0—high, middle, and low—from the same voice. The middle F0 distribution was paired with either the high F0 distribution (rendering it relatively "low") or the low F0 distribution (rendering it relatively "high"). Figure 1 illustrates the approach. This design allowed us to replicate previous findings at a lower F0 range (below 200 Hz) than has been previously investigated. Even more importantly, it established an approach with which to investigate how listeners treat identical acoustic information (stimuli with middle-range F0s) in the context of contextual information that may support selective "binning" of information across acoustic dimensions. In Experiment 1, this contextual information was simply the range of experienced F0 frequencies.

Participants

Assuming a moderate effect size as observed in previously published research employing a highly similar paradigm (Idemaru & Holt, 2011, 2014), at least 15 participants are needed to achieve high power ($1 - \beta = 0.8$, $p < .05$ two-tailed; Faul, Erdfelder, Buchner, & Lang, 2009). We thus tested 30 total participants in Experiment 1 as there were two groups. Participants were undergraduate students (ages 18 to 25 years) from Carnegie Mellon University and University of Pittsburgh and participated for either university credit or a small payment. All participants were native American-English speakers with no exposure to a second language before age two. All reported normal hearing.

Participants were randomly assigned to one of two groups (high F0 range; low F0 range) defined by the F0 × VOT distributions sampled across the experiment. The high F0 range group ($N = 15$) experienced speech sampling 200–300 Hz F0. The low F0 group ($N = 15$) experienced speech from the same voice that sampled 120–220 Hz F0. The protocol in this and subsequent experiments

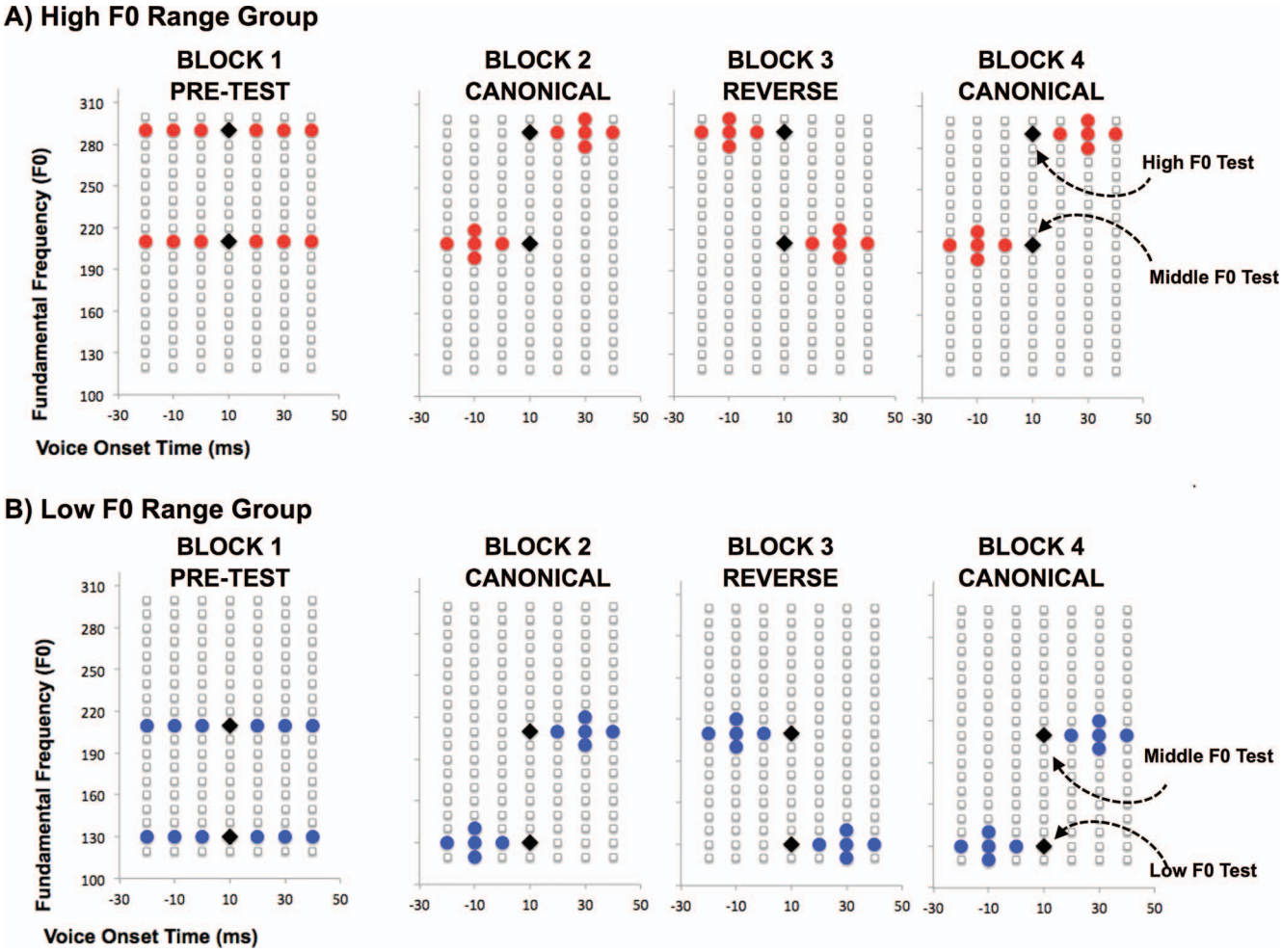

Figure 1. Schematic of Experiment 1 stimuli and design. Each small open square illustrates a single stimulus in the voice onset time (VOT) by fundamental frequency (F0) acoustic space. Stimuli highlighted with large, filled symbols represent those used in Experiment 1. Exposure trials are illustrated as colored circles. Test trials are shown as black diamonds. One group of participants heard stimuli sampling a high F0 range (200–300 Hz, Panel A, filled circles in red). Another group heard stimuli sampling a low F0 range (120–220 Hz, Panel B, filled circles in blue). For each group, the pretest assessed contributions of F0 and VOT to /b/–/p/ categorization without introducing a F0 × VOT correlation in short-term input. The exposure trials of the first experimental block sampled the F0 × VOT space in a manner such that the F0 × VOT correlation was consistent with long-term regularities of English (canonical). The second block introduced an "artificial accent" that reversed this regularity (reverse). In the third block, the canonical English regularity was reestablished. Exposure trials had perceptually unambiguous VOT. Test stimuli possessed perceptually ambiguous VOT and differed in F0. The test stimuli in the high F0 range group had a high (290 Hz) F0 and a middle (210 Hz) F0. The low F0 range group had a middle (210 Hz) and low (130 Hz) F0. See the online article for the color version of this figure.

was approved by the Carnegie Mellon University Institution Review Board.

Stimuli

The stimuli were created from natural productions from a female native speaker of American English who recorded several tokens of the words *beer* and *pier* in a sound attenuated booth (44.1-kHz sampling rate). A single token of *beer* and one of *pier* with similar F0 contours were selected to serve as the natural speech templates for stimulus creation. Manipulations to this pair using Praat (Boersma & Weenink, 2017) resulted in a two-dimensional stimulus space sampling VOT and F0 and varying perceptually from *beer* to *pier*.

VOT was manipulated following the approach of McMurray, Tanenhaus, and Aslin (2002). Fifteen slice points were identified in the *beer* token, with the first occurring 2 ms after the first zero-crossing in the stimulus waveform. The subsequent splice points were positioned in approximately 5-ms increments, with the constraint that the points occurred at zero-crossings. This same

approach was used to identify 15 splice points in the *pier* token. Using the *beer* stimulus as a starting point, the acoustic waveform from a particular splice point was eliminated from *beer* and replaced with the corresponding waveform from the *pier* stimulus. In each case, the substitution was made from the splice point to the end of the waveform. In this way, each stimulus was identical except for the acoustic waveform prior to the splice point. We created a 0-ms VOT token by replacing the burst of the original *beer* token with the burst of the *pier* token. We created the negative VOT durations by inserting a 10-ms (or 20-ms) prevoicing from the *beer* token before the burst of the 0-ms VOT mixture. In all, this resulted in a 15-step series varying perceptually from *pier* to *beer* along an acoustic series corresponding to approximately −20 to 50 ms VOT.

As a manipulation check, 10 undergraduate students who did not participate in any of the main experiments labeled these 15 stimuli as either *beer* or *pier* across 10 repetitions in a randomized order. Based on this, stimulus Step 7 (approximately 10 ms VOT) along the 15-step series was determined to be most perceptually ambiguous (mean 46% /p/ responses, 54% /b/ responses).

Having verified this 10-ms VOT stimulus as perceptually ambiguous, we sampled seven stimuli from the original 15-step VOT series. These stimuli varied from approximately −20 ms to 40 ms VOT in 10-ms steps such that 10 ms VOT was the midpoint of the acoustic speech series. Based on the pilot results, the first three stimulus steps (−20, −10, 0 ms VOT) were expected to be identified as *beer* and the last three stimulus steps (20, 30, 40 ms VOT) were expected to be identified as *pier.* The middle stimulus (10 ms) was perceptually ambiguous, as verified by the pilot results.

We next manipulated the F0 of each stimulus along the VOT series to create a two-dimensional stimulus grid varying across F0 and VOT. We manually manipulated the F0 using Praat such that the onset F0 frequency varied from 120 to 300 Hz in 10-Hz steps for each stimulus along the seven-step VOT series. From this onset F0 frequency, F0 decreased quadratically across the stimulus duration to end 30 Hz lower than onset frequency. In all, this approach resulted in a two-dimensional F0 × VOT acoustic space with the potential to sample 133 unique stimuli as illustrated by the open symbols in Figure 1.

Procedure

On each trial, a single stimulus sampled from the F0 × VOT acoustic space was presented to both ears over headphones and participants responded *beer* or *pier* as quickly and accurately as possible with a key press. Trials were separated by 500 ms.

Modeling the methods of Idemaru and Holt (2011, 2014), the sampling of this two-dimensional acoustic space varied across four blocks, as shown in Figure 1. In the first, the *pretest* block, the relationship between F0 and VOT was balanced to assess the baseline influence of F0 on /b/–/p/ categorization. In the next three blocks, the majority of trials were *exposure* trials (83%, Figure 1, filled circles) designed to provide listeners with short-term experience with a correlation between F0 and VOT that either mirrored the English regularity (canonical block) or introduced an "artificial accent" by reversing it (reverse block). For the exposure stimuli, the principal cue to /b/–/p/ categorization, VOT, signaled /b/–/p/ category membership on every trial.

Each block also included two *test* stimuli (Figure 1, filled diamonds; 17% of trials). Unlike exposure trials, test trials were defined by perceptually *ambiguous* VOT (10 ms VOT, as determined by the pilot experiment) and only F0 varied (see Figure 1). Thus, because VOT is perceptually ambiguous, the extent to which /b/–/p/ categorization differs across test trials serves as a measure of listeners' reliance on F0 in /b/–/p/ categorization, and how it changes as a function of the artificial accent introduced the short-term F0 × VOT regularity experienced across exposure trials within a block.

The participant groups differed in the range of F0 experienced across exposure trials. The high F0 range group (Figure 1a) experienced exposure stimuli sampling 200–300 Hz F0, whereas the low F0 range group (Figure 1b) experienced exposure stimuli sampling 120–220 Hz F0. Across each participant group, one of the two test stimuli had a middle-range F0 (210 Hz). The F0 of the other test stimulus depended on group; the high F0 range condition included a test stimulus with a 290 Hz F0, whereas the low F0 range condition had a test stimulus with a 130 Hz F0. Critical to the present study, the middle-range test stimulus with a 210 Hz F0 was *relatively low* for the high F0 range group and *relatively high* for the low F0 range group.

Pretest. In the pretest block, stimuli sampled the two-dimensional acoustic space defined by F0 and VOT without introducing a short-term correlation between the two acoustic dimensions. As shown in Figure 1, 14 pretest stimuli sampled the F0 × VOT space such that all seven steps across the VOT series (−20 ms to 40 ms in 10-ms steps) were presented at a high and a low F0 (290 and 210 Hz for the high F0 range group; 210 and 130 Hz for the low F0 range group). The exemplars with an F0 of 210 Hz were exactly the same between the two groups. Stimuli were presented across 10 repetitions in random order for a total of 140 pretest trials.

Canonical blocks. In the canonical blocks, the exposure stimuli sampled the acoustic F0 × VOT space in a manner consistent with patterns present in English: short-VOT stimuli had lower F0s whereas long-VOT stimuli had higher F0s. For the high F0 range group, the short VOT values (VOTs −20, −10, and 0 ms) unambiguously signaled /b/ and were paired with the (here relatively *lower*) middle-range F0s (200, 210, and 220 Hz), whereas the long VOT values (20, 30, 40 ms) unambiguously signaled /p/ and were paired with high-range F0s (280, 290, and 300 Hz). For the low-range F0 group, the short VOTs were paired with the low-range F0s (120, 130, and 140 Hz) whereas the longer VOTs were paired with the (here relatively *higher*) middle-range F0s (200, 210, and 220 Hz). For each group, the two test stimuli were also presented in the canonical blocks. Thus, in all there were 12 unique exemplars in the F0 × VOT acoustic space (10 exposure, two test) presented across 10 randomized repetitions for 120 total trials in canonical blocks. Figure 1 illustrates the stimulus sampling for canonical blocks (larger colored circle symbols).

Reverse block. The reverse block was highly similar to the canonical blocks, except that the F0 × VOT correlation across exposure stimuli was opposite that of the canonical blocks. Shorter VOTs (−20, −10, 0 ms) were paired with higher F0s and longer VOTs were paired with lower F0s. Because this relationship runs counter to the long-term regularities of English, the reverse block introduces an "artificial accent" in the short-term speech input. The test stimuli were the same as in the canonical block. In all, there

were 120 trials in the reverse block total (10 exposure stimuli, two test stimuli, 10 repetitions; see Figure 1).

While seated comfortably in a sound-attenuated booth, participants heard the speech stimuli (RMS-matched in overall energy) diotically over headphones (Beyer DT-150, Heilbonn, Germany) with visual indicators of the two response choices (clipart pictures of a beer, a pier) with the spatial layout on a monitor mounted at eye level corresponding to the spatial organization of the keyboard response keys. On each trial of each block, participants simply reported whether they had heard *beer* or *pier* with a key press. Stimulus presentation and response tracking was under the control of E-Prime (Psychology Software Tools, Inc.). A response triggered the next trial after a 500 ms pause. All participants experienced the pretest, a canonical block, a reverse block, and a final canonical block. There was a brief break after each block, but the participants were not informed about any differences among blocks. In total, there were 500 trials across the experiment, which was completed in under 1 hr.

Results

Pretest. We first examined /b/–/p/ categorization in the pretest block, within which there was no F0 × VOT correlation in the short-term input. As shown in Figure 2, a 7 (VOT) × 2 (F0) × 2 (Group) mixed-model analysis of variance (ANOVA) on the mean percentage of *pier* responses in the pretest block revealed a significant main effect of VOT, $F(6, 60) = 179.84$, $p < .001$, $\eta_p^2 = .95$, indicating that the manipulation of stimulus VOT resulted in robust changes in the proportion of *pier* responses. There was also a significant effect of F0, $F(1, 10) = 50.97$, $p < .001$, $\eta_p^2 = .84$, indicating its influence in /b/–/p/ categorization. The participants in different groups did not differ in overall proportion *pier* responses, $F(1, 10) = 1.71$, $p = .220$, $\eta_p^2 = .15$, indicating no overall bias in the proportion *pier* responses as a function of the F0 sampling range.

Among the two-way interactions, only the VOT × F0 interaction was significant, $F(6, 60) = 19.42$, $p < .001$, $\eta_p^2 = .66$; the influence of F0 on /b/–/p/ categorization was impacted by VOT, consistent with prior demonstrations that F0 exerts its effect most robustly when VOT is ambiguous (Abramson & Lisker, 1985; Idemaru & Holt, 2011; Idemaru & Holt, 2014). There was no significant three-way interaction, $F(6, 60) = .54$, $p = .778$, $\eta_p^2 = .05$, indicative of a similar influence of F0 on /b/–/p/ categorization across VOT whether F0 sampled a high or a low frequency range.

We next examined the simple interaction between F0 and group across the stimuli with the most ambiguous VOT (10 ms) that would serve as test stimuli across the canonical and reverse blocks (large symbols in Figure 2). The interaction was not significant, $F(1, 10) = .68$, $p = .430$, $\eta_p^2 = .06$. The influence of F0 on /b/–/p/ categorization was comparable whether the pretest stimuli sampled a high or a low F0 frequency range. The simple effect of F0 (i.e., difference in the proportion of *pier* responses between the high and middle or between middle and low F0s) was significant for both the high F0 range group ($M_{F0 = 290} = .78$, $SE = .09$, 95% confidence interval [CI] [.60, .97]; $M_{F0 = 210} = .37$, $SE = .09$, 95% CI [.17, .56]), $F(1, 10) = 33.78$, $p < .001$, $\eta_p^2 = .87$, and the low F0 range group ($M_{F0 = 210} = .70$, $SE = .08$, 95% CI [.51, .89]; $M_{F0 = 130} = .20$, $SE = .09$, 95% CI [.01, .39]), $F(1, 10) = 48.65$, $p < .001$, $\eta_p^2 = .93$. Participants in each group used F0 to inform /b/–/p/ categorization when VOT was perceptually ambiguous and, as evidenced by the lack of a three-way interaction, did so equivalently across high and low ranges of F0 frequency ranges.

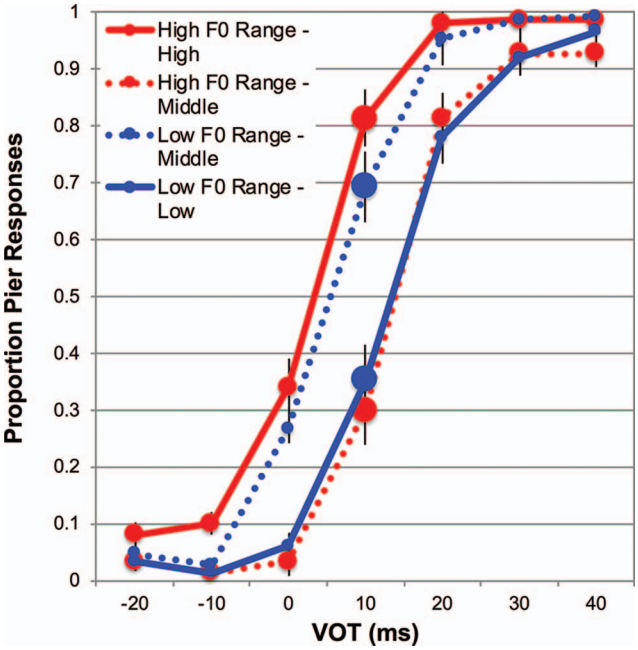

Figure 2. Results of the Experiment 1 pretest. The proportion of *pier* responses are plotted as a function of voice onset time (VOT) and fundamental frequency (F0). For the high F0 range group (in red; lighter grey in print version), the high F0 was 290 Hz and the low F0 was 210 Hz. For the low F0 range group (in blue; darker grey in print version), the high F0 was 210 Hz and the low F0 was 130 Hz. Note that data plotted with dashed lines show the groups' responses to the same stimuli (F0 = 210 Hz) in contexts in which the F0 is relatively high (low F0 range, blue) or relatively low (high F0 range, red). The larger symbols at 10 ms VOT highlight the test stimuli that assess the influence of F0 on /b/–/p/ categorization across the other blocks of the experiment. Error bars reflect standard error of the mean. See the online article for the color version of this figure.

Note from the descriptive statistics that the mean proportion of *pier* responses to the test stimulus with a middle-range F0 (210 Hz) varied across participant groups. Among the high F0 range group, this stimulus was categorized as /p/ on average 37% of the time whereas among the low F0 range group this same stimulus was categorized as /p/ an average of 70% of the time. This difference was significant in an independent samples *t* test, $t(10) = 2.45$, $p = .034$. Participants were less likely to label the test stimulus with a 210 Hz F0 as *pier* in the high F0 range group (for which 210 Hz F0 was relatively low) than in the low F0 range group (for which it was relatively high). The acoustics of this middle F0 test stimulus led to different /b/–/p/ categorization as a function of the range of F0s in which it was experienced.

In all, these data tell us that what counts as "low" versus "high" F0 in informing speech categorization is relative to the range of F0 experienced in the short-term input. This informs the research questions at play in the present study. The distribution of experience across an acoustic dimension can be a source of information across which listeners might "bin" incoming regularities. More broadly, the pretest demonstrates that the sampling of stimuli

within acoustic space provides information for context-dependent speech categorization; we return to this issue in the General Discussion.

Experimental blocks: Exposure stimuli. The exposure stimuli sampled acoustic space such that VOT provided unambiguous acoustic information with which to categorize /b/–/p/. We verified that participants were able to use this information to guide /b/–/p/ categorization responses by examining the mean proportion of *pier* responses to long (/p/-like) versus short (/b/-like) VOTs. Categorization of exposure stimuli based on VOT was highly accurate. Exposure stimuli with long VOT values (20, 30, 40 ms) were most often categorized as /p/. The mean proportion of *pier* responses for long VOT exposure stimuli in the high F0 range group was .97 (SE = .02), .92 (SE = .03), and .96 (SE = .03) for the three experimental blocks (canonical, reverse, canonical), respectively. For the low F0 range group, these values were .95 (SE = .02), .88 (SE = .03) and .94 (SE = .03), respectively.

Likewise, short-VOT exposure stimuli (−20, −10, and 0 ms) were most often categorized as /b/ (with a low proportion of *pier* responses). The mean proportion of *pier* responses to short-VOT exposure stimuli in the high F0 range group was .07 (SE = .03), .12 (SE = .03) and .04 (SE = .02) across the three experimental blocks, respectively. The proportion of *pier* responses for the low F0 range group was .09 (SE = .03), .07 (SE = .03) and .12 (SE = .04) for the three experimental blocks, respectively. This highly accurate exposure stimulus categorization confirmed that VOT robustly signaled English /b/–/p/ categories across blocks.

Experimental blocks: Test stimuli. Categorization of test stimuli with *perceptually ambiguous* VOT served as the primary measure of how short-term regularities conveyed by the exposure stimuli impacted the effectiveness of F0 in signaling /b/–/p/ categories. We first examined the proportion of *pier* responses to test stimuli using a 2 (F0) × 3 (Block) × 2 (Group) ANOVA, whereby the block and F0 factors were manipulated within participants whereas the group factor was manipulated across participants. As Figure 3 shows, there was a significant main effect of F0, $F(1, 10) = 86.39$, $p < .001$, $\eta_p^2 = .90$, and no main effects of either block ($F < 1$) or group, $F(1, 10) = 1.10$, $p = .32$, $\eta_p^2 = .10$.

Most critical to assessing the influence of the introduction of the artificial accent in the reverse block, there was a significant interaction between test stimulus F0 and block, $F(1, 10) = 7.88$, $p = .019$, $\eta_p^2 = .44$, such that the influence of F0 on /b/–/p/ categorization varied across blocks. Additionally, there was a significant interaction between F0 and group, $F(2, 20) = 24.67$, $p < .001$, $\eta_p^2 = .87$, as the overall influence of F0 on /b/–/p/ categorization was influenced by the differential sampling of the F0 range across groups. There was no interaction of block and group ($F < 1$) and no three-way interaction ($F < 1$).

We next examined the simple interaction between F0 and block separately for each group. For the high F0 range group (see Figure 3a), there was a significant main effect of F0, $F(1, 5) = 70.74$, $p < .001$, $\eta_p^2 = .93$, but no effect of block ($F < 1$). There was a significant interaction between F0 and block, $F(2, 10) = 14.73$, $p = .001$, $\eta_p^2 = .75$, indicating the modulation of the influence of F0 on /b/–/p/ categorization as a function of short-term input regularities across blocks. Specifically, the simple effect of F0 was significant in each of the two canonical blocks: Canonical 1, $F(1, 5) = 46.09$, $p = .001$; Canonical 2, $F(1, 5) = 75.63$, $p < .001$, but not in the reverse block, $F(1, 5) = 9.42$, $p = .026 < .017$ (with alpha adjusted for three comparisons). Upon introduction of the artificial accent in the reverse block, F0 was down-weighted in its informativeness to /b/–/p/ categorization.

For the low F0 range group (see Figure 3b), there was a significant main effect for F0, $F(1, 5) = 21.81$, $p = 005$, $\eta_p^2 = .81$, and a marginal effect of block, $F(1, 5) = 3.38$, $p = .076$, $\eta_p^2 = .40$. Critically, the interaction between F0 and block was significant, $F(2, 10) = 11.11$, $p = .003$, $\eta_p^2 = .69$. The simple effect of F0 was significant in each of the canonical blocks: Canonical 1, $F(1, 5) = 30.94$, $p = .003$; Canonical 2, $F(1, 5) = 24.39$, $p = .004$, but not in the reverse block ($F < 1$). In all, these results provide strong evidence for dimension-based statistical learning across both the high and low F0 ranges. Each group of participants down-weighted reliance on F0 in /b/–/p/ categorization upon introduction of the artificial accent that reversed the typical relationship of F0 and VOT in the reverse block.

We next examined how participants responded to the test stimulus with the same absolute F0 value (210 Hz) in different distributions of the local statistics across the experimental blocks as a function of the F0 range sampled with a 3 (Block) × 2 (Group) ANOVA. The main effect of group, $F(1, 10) = 8.70$, $p = .015$, $\eta_p^2 = .47$, and the interaction of group with block were significant, $F(2, 20) = 6.50$, $p = .007$, $\eta_p^2 = .39$. The main effect of block was not significant ($F < 1$). A simple effects analysis revealed that participants in the low F0 range group responded to the 210 Hz F0 test stimulus more often as *pier* than those in the high F0 range group in both canonical blocks: Canonical 1, $F(1, 10) = 16.41$, $p = .002$; Canonical 2, $F(1, 10) = 14.96$, $p = .003$. This group difference was not significant in the reverse block ($F < 1$). This difference in categorization of the 210 Hz F0 test stimuli is consistent with participants' use of *relative* F0 across a distribution of experienced F0s in the high (for which 210 Hz was relatively low) and low (for which 210 Hz was relatively high) F0 range groups.

In sum, the results of Experiment 1 confirm that the rapid learning found in the previous studies of dimension-based statistical learning (Idemaru & Holt, 2011, 2014) can be replicated in a lower F0 range. Even more importantly, the pattern of dimension-based statistical learning observed across the groups is affected by the *relative,* rather than absolute, values experienced across the F0 dimension. Dimension-based statistical learning is impacted not only by the correlation of acoustic dimensions, but also by the range of the experienced dimension.

Experiment 2

The purpose Experiment 2 was to examine whether listeners are able to track overlapping F0 × VOT regularities or whether global acoustic distribution regularities instead predict patterns of categorization. Using the stimuli from Experiment 1 as a starting point, we manipulated voice acoustics to create two separate F0 × VOT acoustic stimulus spaces corresponding to two talkers. The stimulus spaces had identical F0 × VOT information but varied in voice quality conveying talker. We sampled these acoustic spaces in a manner whereby the high-range F0 regularities were sampled with a "female" voice and the low-range F0 regularities were sampled with a "male" voice. This created a scenario in which all

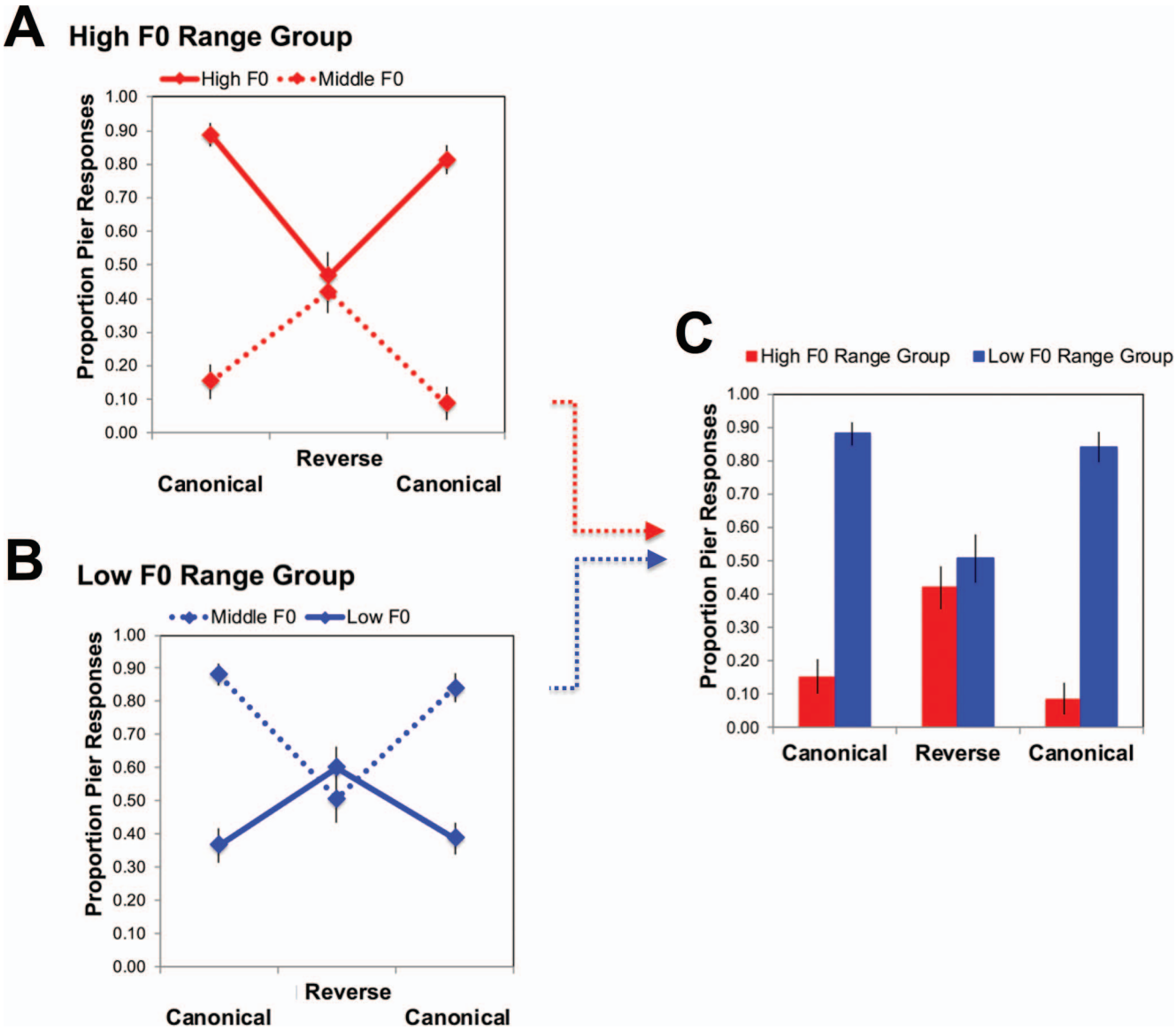

Figure 3. Results of Experiment 1. (A) Mean proportion *pier* responses for the high F0 range group across test stimuli with high (290 Hz) and middle (210 Hz) F0s. (B) Mean proportion of *pier* responses for the low F0 range group across test stimuli with middle (210 Hz) and low (130 Hz) F0s. Note that the middle (210 Hz, plotted with dotted lines) test stimulus was shared across groups. (C) The mean proportion *pier* responses to the middle F0 test stimulus (210 Hz) across the high F0 range group (in red; light grey in print) and the low F0 range group (in blue; dark grey in print). Error bars reflect standard error of the mean. See the online article for the color version of this figure.

listeners experienced the middle-range F0 × VOT distribution as a relatively "low" frequency distribution for the female voice, and also as a relatively "high" frequency distribution for the male voice in the same block of trials. If dimension-based statistical learning is driven exclusively by the accumulation of experience across the acoustic distributions then stimuli sampling a canonical F0 × VOT correlation for one voice and a reverse F0 × VOT correlation for the other voice should produce no F0 down-weighting because the opposing distributions eliminate any F0 × VOT correlation in short-term input. Alternatively, if voice information is sufficient to allow listeners to track the distinct F0 × VOT regularities co-evolving across voices, then F0 down-weighting should be evident across the voice conveying an artificial accent even as F0 continues to signal /b/–/p/ categorization for the voice sampling the canonical English F0 × VOT correlation.

Method

Participants. Assuming a moderate effect size as observed in previously published research employing a highly similar para-

digm (Idemaru & Holt, 2011, 2014), at least 15 participants are needed to achieve high power ($1 - \beta = 0.8$, $p < .05$ two-tailed, Faul et al., 2009). Inasmuch as the present design involves a potentially subtler effect than tested previously, we aimed to test 30 total participants in Experiment 2. Here, we report the full data sets acquired from 28 undergraduate students (ages 18 to 25 years). Students were from Carnegie Mellon University and University of Pittsburgh and participated in this experiment for either university credit or a small payment. All participants were native American English speakers with no exposure to a second language before the age of two. All reported normal hearing. None of them had participated in the previous experiment.

Stimuli. The stimuli were based on the two-dimensional F0 × VOT acoustic space created for Experiment 1. For Experiment 2, two new acoustic spaces were created using the automated "change gender" function in Praat. This function shifts formant frequencies as a ratio of the original sound via manipulation of sampling frequency. The manipulation shifted the female formant frequencies in the original Experiment 1 speech tokens toward a more exaggerated female voice (Praat ratio of 1.1) or toward a more male voice (Praat ratio of 0.9). The F0 of the stimuli was not affected by this manipulation, but the formant shift resulted in a distinct change in voice quality, readily apparent to listeners. For simplicity in verbal description, we refer to the stimuli shifted toward higher formant frequencies as "female" and those shifted toward lower formant frequencies as "male." For our purposes it is not important that listeners identify these voices as female and male, specifically. It is only important that they be able to discriminate the voices as distinct talkers.

To verify that listeners hear the stimulus sets as a qualitative shift in talker, we conducted a brief pilot study in which a separate group of listeners ($N = 15$) drawn from the same sample as the study participants discriminated the middle F0 (210 Hz) test stimulus in each voice. In an AX discrimination task, participants heard two repetitions of the middle F0 test stimulus in the "female" voice, two repetitions of the same token in the "male" voice, as well as female–male and male–female pairings. Listeners indicated whether the voice was the same or different on each of 20 total trials (10 same, 10 different). Performance was highly accurate ($M = 95\%$ accuracy, $SE = .41$), verifying the discriminability of the voices.

Procedure. The procedure was identical to that of Experiment 1, except in the manner that stimuli were sampled from the F0 × VOT acoustic space. Trials sampling the high F0 range (now in a female voice) and low F0 range (in a male voice) were mixed within a block. All participants experienced the same stimulus sampling across the female and male voices in a within-participant design.

Pretest. In the pretest, the seven stimuli across the VOT series were presented at two F0s for each voice. For the female voice, the F0s were 210 and 290 Hz, sampling the higher F0 range. For the male voice, the F0s were 130 and 210 Hz, sampling the lower F0 range (see Figure 4). There were 10 repetitions of each stimulus for a total of 140 stimuli per block, with separate pretest blocks for female and male voices counterbalanced in order. The stimulus sampling in F0 × VOT space for the pretest was like that shown in Figure 1 for the Experiment 1 pretest.

Experimental (canonical, reverse) blocks. The female and male voices were mixed in presentation in the experimental blocks, as illustrated in Figure 4. Stimuli were sampled from the higher F0 range (200–300 Hz) for the female voice whereas stimuli were sampled from the lower F0 range (120–220 Hz) for the male voice. Thus, listeners experienced the F0 × VOT distribution sampling the middle F0 distribution in each voice. This distribution was a relatively low F0 for the female voice, but a relatively high F0 for the male voice. Yet, these stimuli shared identical F0 and VOT information.

Figure 4 illustrates the sampling of stimuli across Experiment 2 experimental blocks. In one block, stimuli were sampled such that both female and male voice exposure stimuli (solid circles, Figure 4) possessed the canonical English F0 × VOT correlation ($female_{CAN}$–$male_{CAN}$). In another block, exposure stimuli sampled a reverse F0 × VOT correlation for the female voice and a canonical F0 × VOT correlation for the male voice ($female_{REV}$–$male_{CAN}$). An additional block involved exposure stimuli sampling the canonical correlation for the female voice and the reverse correlation for the male voice ($female_{CAN}$–$male_{REV}$). Across each block, there were four test stimuli with perceptually ambiguous VOT (10 ms, solid diamonds Figure 4); two stimuli in the female voice sampled the higher F0 range (210 and 290 Hz) and two stimuli in the male voice sampled the lower F0 range (130 and 210 Hz). Thus, for each voice there were 10 exposure stimuli and two test stimuli (12/voice, 24 total stimuli). Each was presented 10 times per block for a total of 240 trials. These blocks immediately followed the two pretest blocks. The $female_{CAN}$–$male_{CAN}$ was always presented first, with the remaining two experimental blocks counterbalanced in order across participants to balance presentation of the artificial accent (reverse block) in the male and female voices. Throughout the experiment, the task was simply to identify stimuli as *beer* versus *pier*. As in Experiment 1, there were short breaks between blocks and participants were not informed about the changing relationships of acoustic dimensions across blocks. In all, there were 1,000 trials.

Results

Pretest. We first evaluated /b/–/p/ categorization at pretest with a 7 (VOT) × 2 (F0) × 2 (Voice) ANOVA on the mean proportion of *pier* responses in the pretest blocks, with all factors manipulated within participants. Figure 5 shows the results. There were robust main effects of both VOT, $F(6, 162) = 562.59$, $p < .001$, $\eta_p^2 = .95$, and F0, $F(1, 27) = 87.51$, $p < .001$, $\eta_p^2 = .76$, indicating that each contributed to /b/–/p/ categorization. There was no main effect of voice ($F < 1$), indicating no bias toward greater *pier* responses for one voice versus the other. The two-way interactions between VOT and F0, $F(6, 162) = 17.72$, $p < .001$, $\eta_p^2 = .40$, VOT and talker, $F(6, 162) = 10.30$ $p < .001$, $\eta_p^2 = .28$, and F0 and talker, $F(1, 27) = 73.44$, $p < .001$, $\eta_p^2 = .73$, were all significant. There was also a significant three-way interaction, $F(6, 162) = 17.31$, $p < .001$, $\eta_p^2 = .39$, indicating that the influence of F0 on /b/–/p/ categorization was modulated by VOT and that this modulation was impacted by voice.

We next conducted a planned test of the simple interaction between F0 and voice across the stimuli that would serve as test stimuli with perceptually ambiguous VOT (10 ms) in the experimental blocks (Figure 5, large symbols). The interaction was significant, $F(1, 27) = 57.69$, $p < .001$, $\eta_p^2 = .68$, indicating that

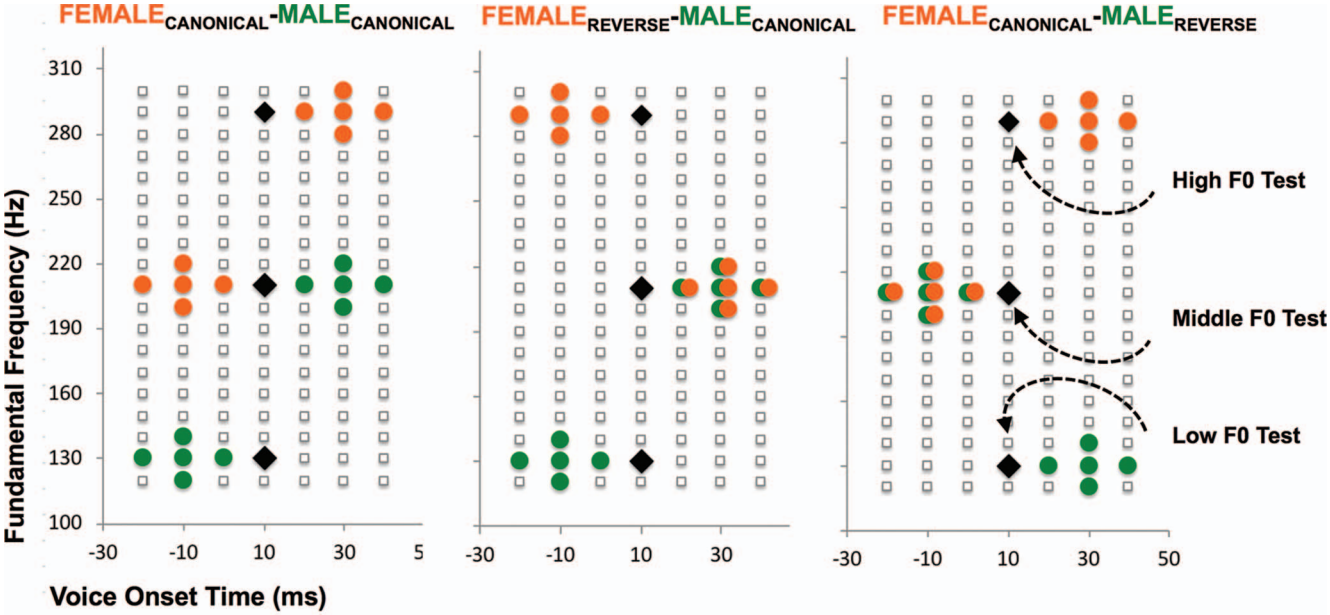

Figure 4. Experiment 2 and Experiment 3 stimuli. Schematic illustration of the sampling of the voice onset time (VOT, in ms) by fundamental frequency (F0, in Hz) acoustic stimulus space for Experiments 2 and 3. The stimuli plotted in solid large symbols were presented during the experiment. Large colored circles correspond to exposure stimuli; large solid diamonds are test stimuli with perceptually ambiguous VOT. As plotted here, the sampling of exposure stimuli is such that stimuli in the high F0 range sample canonical–reverse–canonical (orange symbols; light grey in print). Those in the low F0 range sample canonical–canonical–reverse (green symbols; dark grey in print). In Experiment 2, the stimuli plotted with orange symbols were synthesized to have a female voice quality and those plotted with green symbols were synthesized to have a male voice quality. In Experiment 3, all stimuli had the same voice quality (equivalent to that of Experiment 1). The high F0 range stimuli, plotted in orange, were paired with a video of a female articulating face. The low F0 range stimuli, plotted in green, were paired with a video of a male articulating face. See the online article for the color version of this figure.

the influence of F0 differed across the female and male voices at pretest. Specifically, there was a robust F0 effect only for the female voice, $F(1, 27) = 136.10$, $p < .001$. The female test stimulus with the high F0 ($M = .73$, $SE = .03$, 95% CI [.67, .80]) was more often categorized as /p/ than the test stimulus with the middle F0 ($M = .21$, $SE = .05$, 95% CI [.12, .30]). Quite surprising in light of the Experiment 1 results, F0 did not influence categorization of the test stimuli for the male voice ($F < 1$; F0 = 210 Hz, $M = .50$, $SE = .05$, 95% CI [.40, .61]; F0 = 130 Hz, $M = .48$, $SE = .05$, 95% CI [.38, .58]). We return to this issue in the General Discussion.

In line with the results of Experiment 1, there was a significant difference in categorization of the middle F0 test stimulus (F0 = 210 Hz) across the female voice sampling the higher F0 range and the male voice sampling the lower F0 range, $t(27) = 5.64$, $p < .001$, with the stimulus categorized more often as *pier* when it was a relatively high F0 in the male voice ($M = .50$, $SE = .05$, 95% CI [.40, .61]) compared to when it was a relatively low F0 in the female voice ($M = .21$, $SE = .05$, 95% CI [.12, .30]). As in Experiment 1, participants treated the middle-range F0 (dashed lines in Figure 5) differently as a function of the range of F0 experienced in the short-term input.

Experimental blocks: Exposure stimuli. As in Experiment 1, we first examined the responses to the exposure stimuli in the experimental blocks to confirm that participants used the perceptually unambiguous VOT to guide categorization of exposure stimuli. For exposure stimuli with relatively longer VOTs (20, 30, 40 ms), the mean proportions *pier* responses were .97 ($SE = .01$), .93 ($SE = .02$), and .90 ($SE = .02$) for the female$_{CAN}$–male$_{CAN}$, female$_{REV}$–male$_{CAN}$, and female$_{CAN}$–male$_{REV}$ blocks, respectively. Across exposure stimuli with shorter VOTs consistent with *beer* (−20, −10, 0 ms), the mean proportion of *pier* responses was .05 ($SE = .01$), .07 ($SE = .02$), and .15 ($SE = .02$) across these same blocks. This confirms highly accurate /b/–/p/ categorization of exposure stimuli on the basis of the perceptually unambiguous VOT.

Experimental blocks: Test stimuli. We next examined categorization of the test stimuli to investigate whether listeners use identical F0 × VOT distributions distinctly across voice. Figure 6 shows the results. A 2 (F0) × 3 (Block) × 2 (Voice) ANOVA, with all factors manipulated within participants revealed a significant main effect of F0, $F(1, 27) = 40.31$, $p < .001$, $\eta_p^2 = .60$, indicating that F0 impacted /b/–/p/ categorization of test stimuli. There was also a main effect of voice, $F(1, 27) = 16.40$, $p < .001$, $\eta_p^2 = .38$, consistent with a greater proportion of *pier* responses for the female voice compared to the male voice. There was no main effect of block ($F < 1$). There was a robust interaction of F0 and block, $F(2, 54) = 51.28$, $p < .001$, $\eta_p^2 = .66$, consistent with

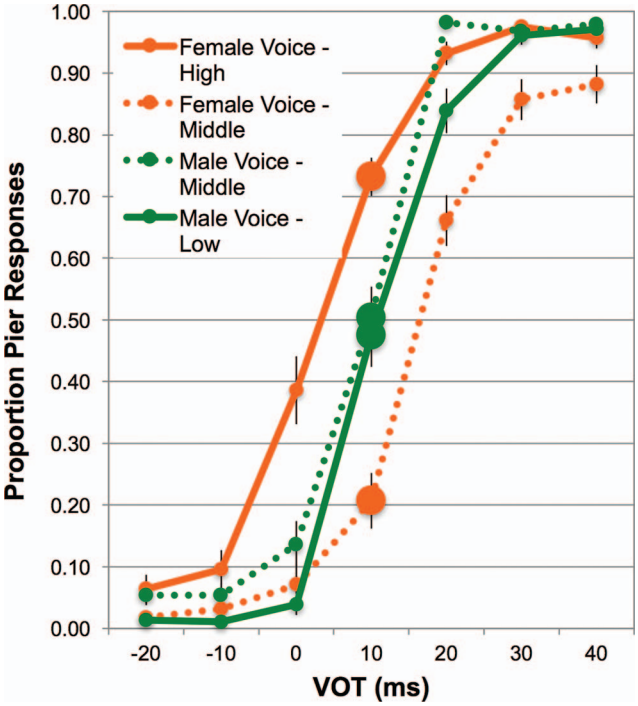

Figure 5. Results of the Experiment 2 pretest. The proportion of *pier* responses are plotted as a function of voice onset time (VOT) as a function of fundamental frequency (F0). For the high F0 range group presented in a female voice (in orange; light grey in print), the high F0 test stimulus was 290 Hz, and the low F0 test stimulus was 210 Hz. For the low F0 range group presented in a male voice (in green; dark grey in print), the high F0 test stimulus was 210 Hz and the low F0 test stimulus was 130 Hz. Note that data plotted with dashed lines show responses to the same stimuli (F0 = 210 Hz) in contexts in which the F0 is relatively high (low F0 range, green) or relatively low (high F0 range, orange). The larger symbols at 10 ms VOT highlight the test stimuli that assess the influence of F0 on /b/–/p/ categorization across the other blocks of the experiment. Error bars reflect standard error of the mean. See the online article for the color version of this figure.

modulation of the effectiveness of F0 in signaling /b/–/p/ categorization as a function of the short-term F0 × VOT regularities across block. There was also an interaction of F0 and voice, $F(1, 27) = 76.09$, $p < .001$, $\eta_p^2 = .74$, indicating different patterns of F0 down-weighting across voice. As shown in Figure 6, this is due to the different short-term regularities present across voice mixed within a block (e.g., canonical–reverse–canonical for the female voice; canonical–canonical–reverse for the male voice). Neither the Block × Voice interaction, $F(2, 54) = 1.16$, $p = .323$, $\eta_p^2 = .04$, nor the three-way interaction, $F(2, 54) = 1.05$, $p = .358$, $\eta_p^2 = .04$, was significant.

We next examined the simple interaction between F0 and block for each voice independently, by testing whether the influence of F0 was modulated by short-term regularities experienced in each voice. For the female voice, there was a significant main effect of F0, $F(1, 27) = 125.79$, $p < .001$, $\eta_p^2 = .82$, and no effect of block ($F < 1$). Most critically, as is apparent in Figure 6a, there was a robust interaction of F0 and block, $F(2, 54) = 22.54$, $p < .001$, $\eta_p^2 = .46$. F0 impacted test trial categorization in each of the blocks: $female_{CAN}$–$male_{CAN}$, $F(1, 27) = 125.33$, $p < .001$; $female_{REV}$–$male_{CAN}$, $F(1, 27) = 10.21$, $p = .004$; $female_{CAN}$–$male_{REV}$, $F(1, 27) = 168.00$, $p < .001$. But, the magnitude of this influence was modulated by how the female voice sampled the F0 × VOT correlation across blocks. The influence of F0 on test stimulus categorization was significantly greater when the female voice sampled the canonical F0 × VOT regularity ($female_{CAN}$–$male_{CAN}$, $M = .50$, $SE = .05$, 95% CI [.41, .60] and $female_{CAN}$–$male_{REV}$, $M = .52$, $SE = .04$, 95% CI [.44, .60]) than when it conveyed the artificial accent ($female_{REV}$–$male_{CAN}$, $M = .19$, $SE = .06$, 95% CI [.07, .31], $ps < .001$). The influence of F0 on /b/–/p/ categorization was down-weighted when the F0 × VOT correlation was specifically reversed for the female voice, even as the male voice maintained the canonical correlation.

The influence of F0 on /b/–/p/ categorization across the male voice was unexpected. Whereas F0 had a robust influence on /b/–/p/ categorization for the low F0 range stimuli in Experiment 1, F0 did not impact categorization of stimuli sampling the same distribution of F0 × VOT space presented in a male *voice* in Experiment 2. Mirroring the Experiment 2 pretest results, there was no significant main effect of F0 ($F < 1$) on categorization of the test stimuli. Nonetheless, there was a significant interaction of F0 and block, $F(2, 54) = 31.58$, $p < .001$, $\eta_p^2 = .54$. Specifically, the F0 effect was significant in the $female_{CAN}$–$male_{CAN}$, $F(1, 27) = 8.97$, $p = .006$, and $female_{CAN}$–$male_{REV}$, $F(1, 27) = 29.99$, $p < .001$, blocks. But, the F0 effect was flipped relative to the canonical English pattern in the latter block (see Figure 6d). Introduction of the artificial accent in the male voice led participants to use F0 in a manner that mirrored the short-term input statistics of the F0 × VOT correlation reversal, rather than merely down-weighting F0 as observed for the female voice (and prior studies; Idemaru & Holt, 2011, 2014). The F0 effect was not significant in the $female_{REV}$–$male_{CAN}$ block, $F(1, 27) = 2.01$, $p = .168$, and the effect size of F0 (i.e., the mean difference scores in the proportion of *pier* responses between the middle vs. low F0 test stimuli for the male speaker) did not differ significantly between the $female_{CAN}$–$male_{CAN}$ ($M = .15$, $SE = .05$, 95% CI [.05, .26]) and the $female_{REV}$-$male_{CAN}$ ($M = .08$, $SE = .05$, 95% CI [–.04, .20]) blocks ($p = .443$), which were both different from the $female_{CAN}$-$male_{REV}$ block ($M = -.27$, $SE = .05$, 95% CI [–.37, –.17]; $ps < .001$). Figure 6c summarizes the influence of F0 on categorization of test stimuli across female and male voices. There was no effect of block ($F < 1$). We return to discuss the unanticipated patterns for the male voice in the General Discussion.

We also examined whether participants responded differently to the test stimuli with the same absolute F0 value (210 Hz) for different speakers with a 3 (Block) × 2 (Voice) ANOVA conducted for categorization of the middle F0 test stimuli (210 Hz) between the two speakers across blocks. The main effect was significant for block, $F(2, 54) = 14.42$, $p < .001$, $\eta_p^2 = .35$, but not for voice, $F(1, 27) = 2.21$, $p = .149$, $\eta_p^2 = .07$. The interaction between block and voice was significant, $F(2, 54) = 8.52$, $p = .001$, $\eta_p^2 = .24$. As can been seen in Figure 6c, the simple effects analysis confirmed that the same test stimulus was categorized more as *pier* for the male speaker than for the female speaker in the $female_{CAN}$–$male_{CAN}$ block, $F(1, 27) = 12.98$, $p = .001$. There was no difference between the two speakers in the other two blocks, $ps > .700$.

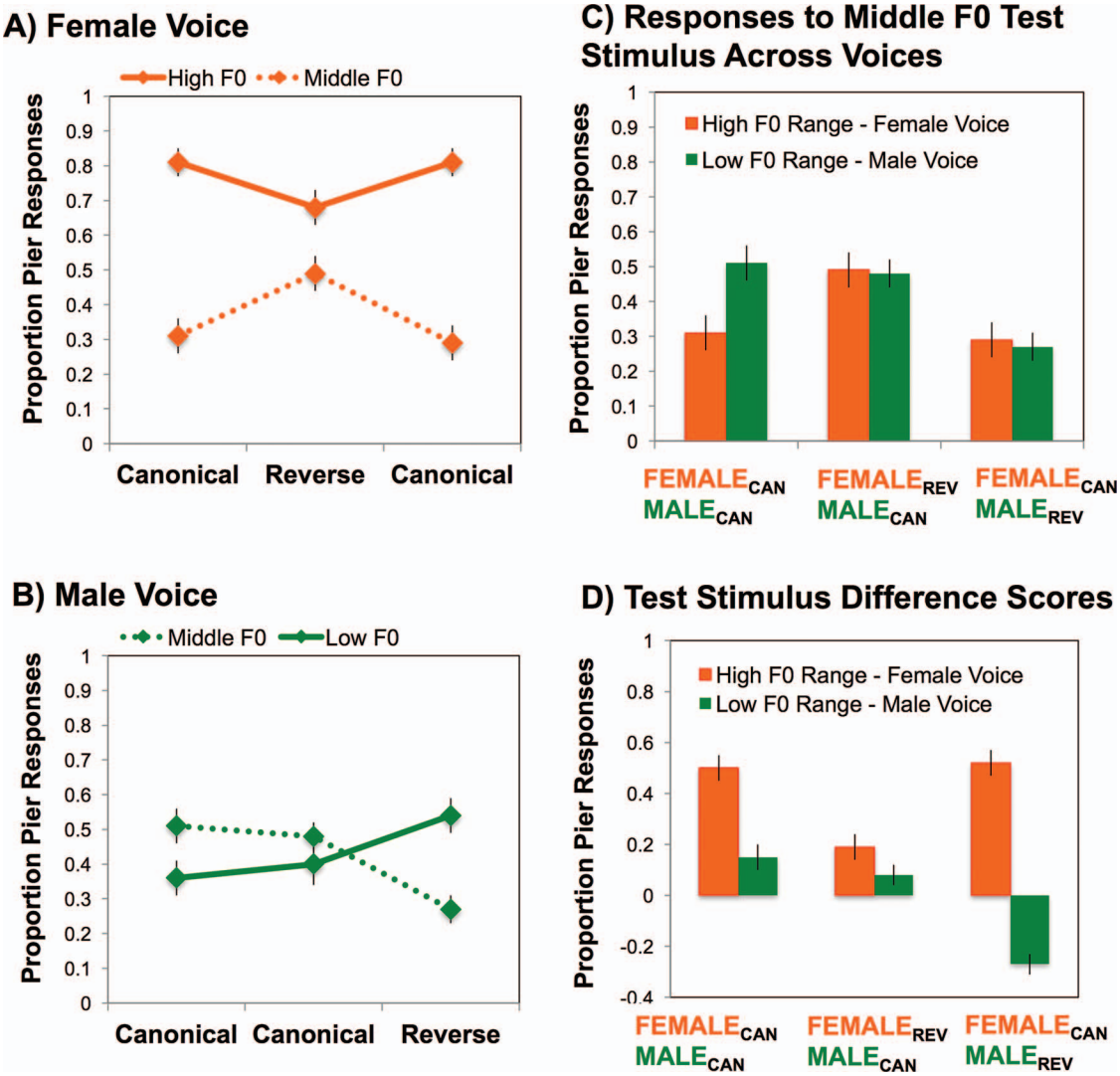

Figure 6. Results of Experiment 2. Mean proportion *pier* responses to Experiment 2 test stimuli as a function of voice quality of the talker (female, male) and the short-term regularity of the exposure stimuli (canonical [CAN], reverse [REV]). (A) Proportion of *pier* responses to the high (290 Hz) and middle (210 Hz) F0 test stimuli as a function of the short-term regularity experienced in the female voice in the high F0 range. (B) Proportion of *pier* responses to the middle (210 Hz) and low (130 Hz) F0 test stimuli as a function of the short-term regularity experienced in the male voice in the low F0 range. Note that the dashed lines in Panels (A) and (B) correspond to responses to the middle F0 test stimulus with the same F0 and VOT characteristics. (C) Mean proportion of *pier* responses to the middle F0 (210 Hz) test stimulus for the female and male voices. (D) Summary of the contribution of F0 to *beer–pier* categorization responses across conditions, plotted as the difference score between the two test stimuli in each condition. Error bars reflect standard error of the mean. See the online article for the color version of this figure.

In sum, dimension-based statistical learning is not the mere accumulation of acoustic distributions. The modulation of the influence of F0 on /b/–/p/ categorization upon introduction of an artificial accent that reverses the F0 × VOT correlation demonstrates that listeners can track distributional regularities across acoustic dimensions in a voice-dependent manner. Because stimuli sampling the male and female voices were intermixed across the experimental blocks of Experiment 2, the global F0 × VOT correlation within the female$_{REV}$–male$_{CAN}$ and female$_{CAN}$–male$_{REV}$ blocks was neutral. Yet, the influence of F0 on /b/–/p/ categorization was modulated in a manner that suggests listeners were able to bin overlapping experience across F0 and VOT acoustic dimensions according to talker-specific voice information. Participants down-weighted reliance on F0 in categorizing the female voice test stimuli in the context of a block in which the F0 × VOT relationship was reversed for the female voice, but canonical for the male voice. Similarly, participants reweighted F0 in categorizing male test stimuli in

response to experiencing the reversed correlation of the artificial accent in the male voice even as they simultaneously relied upon F0 in the typical manner in categorizing female test stimuli. Indeed, in this case listeners' use of F0 came to mirror the reverse F0 × VOT correlation of the artificial accent. Additionally, listeners were able to use the middle-range F0 test stimuli with identical F0 × VOT characteristics differentially across voice. In all, this indicates that dimension-based statistical learning is not driven exclusively by the sampling of acoustic F0 × VOT. Instead, information available from voice quality is used to partition experience across these acoustic dimensions such that listeners can track distinct F0 × VOT regularities across voices. In the most extreme example, listeners relied upon the canonical F0 × VOT mapping for the female voice even as they reversed the mapping in response to an artificial accent presented in the male voice.

Experiment 3

Experiment 2 demonstrated that the perceptual system is able to use additional information, such as voice quality, to track overlapping acoustic distributions in a manner that effectively "bins" talker-specific short-term regularities across acoustic dimensions. Experiment 3 examined this further by testing whether the system can track overlapping acoustic distributions even across *identical* acoustic input. In Experiment 3, the stimuli sampling F0 × VOT acoustic space had a single, consistent voice quality equivalent to that of Experiment 1. Within an Experiment 3 block, acoustic stimuli sampling the high F0 range were paired with a silent video of an articulating female face and acoustic stimuli sampling the low F0 range were paired with a silent video of an articulating male face. This eliminated the distinctive acoustic formant frequency information creating two voices in Experiment 2. Thus, in Experiment 3 the acoustic information signals the same talker. This may encourage "binning" of exemplars experienced across the F0 × VOT space according to the global block-level F0 × VOT distributional regularity, which conveys a neutral F0 × VOT relationship (because the short-term regularities include both canonical and reverse sampling). If listeners are sensitive only to acoustic information, the presence of an artificial accent (reverse sampling) should not lead to down-weighting of the effectiveness of F0 in signaling /b/–/p/ categories. However, if listeners can use the visual information paired with speech exemplars sampling the high F0 range and low F0 range acoustic signal regularities across F0 and VOT, then patterns of dimension-based statistical learning should track with the short-term regularities associated with the videos.

Method

Participants. Assuming a moderate effect size as observed in previously published research employing a highly similar paradigm (Idemaru & Holt, 2011, 2014), at least 15 participants are needed to achieve high power ($1 - \beta = 0.8$, $p < .05$ two-tailed, Faul et al., 2009). Inasmuch as the present design involves a potentially subtler effect than tested previously, we aimed to test 30 total participants in Experiment 3. Here, we report the full data sets acquired from 25 undergraduate students (ages 18 to 25 years). Students were from Carnegie Mellon University and University of Pittsburgh and participated in this experiment for either university credit or a small payment. All participants were native American English speakers with no exposure to a second language before the age of two. All reported normal hearing. None of them had participated in the previous experiment.

Stimuli. The auditory stimuli were identical to those of Experiment 1, with all speech tokens from a single voice. The sampling of the acoustic F0 × VOT space was identical to the sampling of Experiment 2, as shown in Figure 3. The high F0 range stimuli (orange symbols in Figure 4) were differentiated from the low F0 range stimuli (green symbols in Figure 4) by pairing the acoustic stimuli with a video of a female articulating face or a male articulating face, respectively.

Short videos of a female and a male talker (both young Caucasian adults) were recorded by asking each talker to listen to the recording of the *beer–pier* stimulus with the most ambiguous VOT and to repeat the word several times, imitating the rate of speech as closely as possible (see Figure 7 for screenshots). A video clip with the best temporal synchrony with the acoustic stimulus was chosen for each talker and edited using Adobe Premiere. The onset of the word recorded by the camera's built-in microphone was identified, and the utterance recorded in the video was removed and replaced with acoustic stimuli sampling the F0 × VOT acoustic space created for Experiment 1. The acoustic exposure and test stimuli sampling the higher F0 distribution (280–300 Hz) were

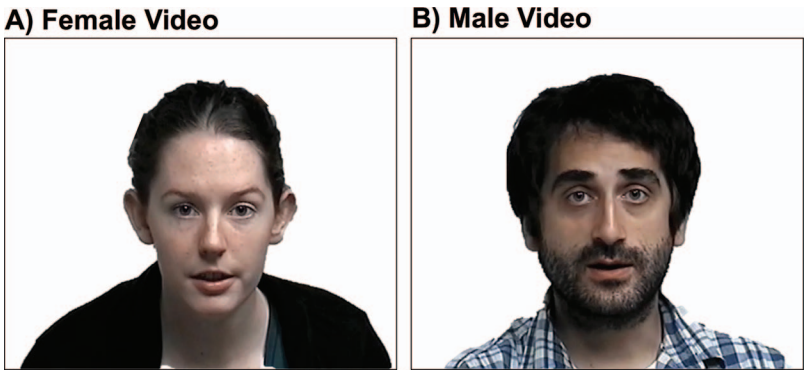

Figure 7. Screenshots of videos from Experiment 3. Panel (A) shows a frame from the female video. Panel (B) shows a frame from the male video. See the online article for the color version of this figure.

dubbed onto the female video whereas the acoustic exposure and test stimuli sampling the lower F0 distribution (120–140 Hz) were dubbed onto the male video. The acoustic exposure and test stimuli sampling the middle F0 distribution (200–220 Hz) were dubbed onto both the female and the male videos. Thus, the acoustic stimuli were identical to those used of Experiment 1; they sampled the F0 × VOT acoustic space in a natural female voice. The visual information conveyed by the single female video and single male video was identical across the acoustic exemplars and so it provided no specific information supporting /b/–/p/ categorization.

Procedure. The procedure followed the approach of Experiment 2 illustrated in Figure 4. In two pretests, trials were blocked by the female and male videos such that participants responded to all seven VOT exemplars at each of the two F0 frequencies associated with the female or male video, respectively (following the pretest acoustic stimulus sampling shown in Figure 1). High F0 range acoustic stimuli were paired with the female video (Figure 4, orange symbols); low F0 range acoustic stimuli were paired with the male video (Figure 4, green symbols). In the Experimental blocks, trials with the female and male videos were mixed together, with the F0 × VOT correlation varying across blocks as in Experiment 2 ($female_{CAN}$–$male_{CAN}$, $female_{REV}$–$male_{CAN}$, and $female_{CAN}$–$male_{REV}$; see Figure 4). The order of the two pretest blocks was counterbalanced across participants, as was the order of the last two Experimental blocks. In all, there were 1,000 trials.

Results

Pretest. We first examined /b/–/p/ categorization at pretest, for which "talkers"—the differential acoustic sampling across the F0 × VOT space (high/low F0 range) and the associated pairing with videos (female/male)—were presented in two separate blocks. A 7 VOT × 2 F0 × 2 talker repeated-measures ANOVA was conducted on the mean proportion *pier* responses. As evident in Figure 8, there were robust main effects of VOT, $F(6, 144) = 527.34$, $p < .001$, $\eta_p^2 = .96$, and F0, $F(1, 24) = 30.68$, $p < .001$, $\eta_p^2 = .56$, indicating the contribution of each acoustic dimension to /b/–/p/ categorization. There was also a significant main effect of talker, $F(1, 24) = 11.73$, $p = .002$, $\eta_p^2 = .33$, indicating a significant shift in proportion *pier* responses as a function of the range of F0 and the video with which these stimuli were paired. The two-way interactions between VOT and F0, $F(6, 144) = 30.56$, $p < .001$, $\eta_p^2 = .56$, VOT and talker, $F(6, 144) = 22.31$ $p < .001$, $\eta_p^2 = .48$, and F0 and talker, $F(1, 24) = 21.19$, $p < .001$, $\eta_p^2 = .47$, were all significant. There was also a significant three-way interaction, $F(6, 144) = 18.01$, $p < .001$, $\eta_p^2 = .43$, indicating that the influence of F0 on /b/–/p/ categorization was modulated by VOT as well as by the talker implicated by the different videos and F0 sampling ranges.

We next conducted a planned test of the simple interaction between F0 and implied talker across the stimuli that would serve as test stimuli with perceptually ambiguous VOT (10 ms) in the experimental blocks (Figure 8, large symbols). This revealed a significant interaction, $F(1, 24) = 37.96$, $p < .001$, $\eta_p^2 = .61$, indicating that the influence of F0 on /b/–/p/ categorization differed across the implied talkers. Specifically, when F0 was sampled from a higher range and paired with the female video, the test stimulus with a high-range F0 (290 Hz) was more likely to be categorized as *pier* ($M = .62$, $SE = .05$, 95% CI [.52, .72]) than the test stimulus with the middle-range F0 (210 Hz; $M = .09$, $SE = .02$, 95% CI [.04, .14]). When F0 was sampled from a lower range and paired with the male video, the test stimulus with the middle-range F0 (210 Hz) was more likely to be categorized as *pier* ($M = .39$, $SE = .06$, 95% CI [.27, .51]) than the low-range F0 (130 Hz; $M = .29$, $SE = .04$, 95% CI [.21, .37]). As is clear from these descriptive statistics, the influence of F0 was stronger for the higher range of F0 paired with the female video than for the lower range F0 paired with the male video although each was significant, $F(1, 24) = 105.26$, $p < .001$ and $F(1, 24) = 4.36$, $p = .048$, respectively.

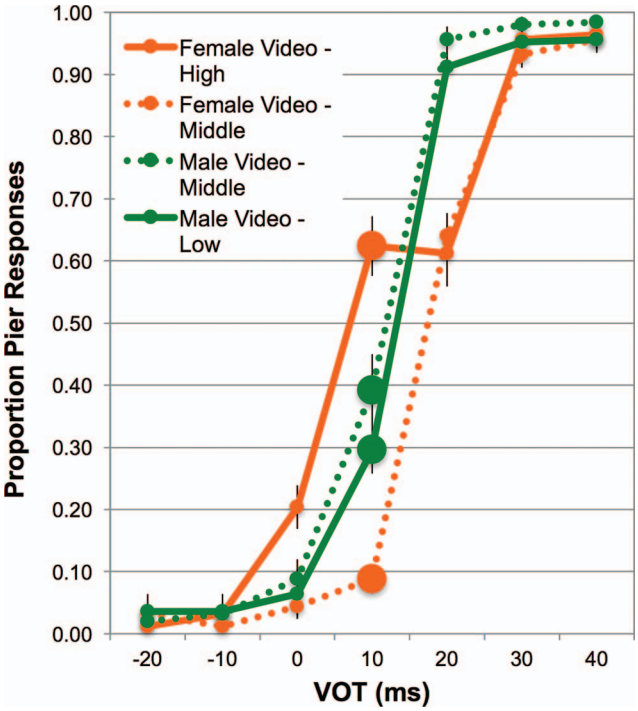

Figure 8. Results of the Experiment 3 pretest. The proportion of *pier* responses are plotted as a function of voice onset time (VOT) as a function of fundamental frequency (F0). For the high F0 range group paired with a video of an articulating female face (in orange; light grey in print), the high F0 test stimulus was 290 Hz and the low F0 test stimulus was 210 Hz. For the low F0 range group paired with an articulating male face (in green; dark grey in print), the high F0 test stimulus was 210 Hz and the low F0 test stimulus was 130 Hz. Note that data plotted with dashed lines show responses to the same stimuli (F0 = 210 Hz) in contexts in which the F0 is relatively high (male/low F0 range, green) or relatively low (female/high F0 range, orange). The larger symbols at 10 ms VOT highlight the test stimuli that assess the influence of F0 on /b/–/p/ categorization across the other blocks of the experiment. Error bars reflect standard error of the mean. See the online article for the color version of this figure.

As in Experiments 1 and 2, there was a significant difference in categorization of the middle F0 test stimulus (F0 = 210 Hz) across the higher F0 range paired with a female video and the lower F0 range paired with a male video, $t(24) = 6.06$, $p < .001$. When the middle F0 test stimulus was presented in the context of a block sampling the low F0 range paired with the male video, it was more often categorized as *pier* ($M = .39$, $SE = .06$, 95% CI [.27, .51]) than when this same stimulus was presented in the context of a

block sampling the high F0 range paired with the female video (M = .09, SE = .02, 95% CI [.04, .14]). The same test stimulus was treated as a relatively low F0 in the context of a higher range of F0s paired with a female video and as relatively high F0 paired with a lower range of F0s paired with a male video even though the acoustic F0 information was identical.[1]

Experimental blocks: Exposure stimuli. We next examined the accuracy of categorization across exposure trials, for which VOT provided perceptually unambiguous information to signal /b/–/p/ categories. Exposure stimuli with longer VOTs (20, 30, 40 ms) consistent with /p/ were categorized consistently as *pier* across the three Experimental blocks (mean proportion *pier* responses; .94 [SE = .01], .93 [SE = .01], and .92 [SE = .01] for the female$_{CAN}$–male$_{CAN}$, female$_{REV}$–male$_{CAN}$, and female$_{CAN}$–male$_{REV}$ blocks, respectively). Exposure stimuli with shorter VOTs (−20, −10, 0 ms) consistent with /b/ were categorized much less often as *pier* (.02 [SE = .01], .04 [SE = .01], and .02 [SE = .01] for the blocks). As in the previous two experiments, categorization performance of exposure stimuli based on VOT was highly accurate. This provided information about how F0 covaried with VOT and /b/–/p/ category membership across the short-term regularities manipulated across blocks.

Experimental blocks: Test stimuli. The mean proportion of *pier* responses to test stimuli across conditions are plotted in Figure 9. As in the prior experiments, categorization of the test stimuli reveals how short-term experience with the short-term regularities in F0 × VOT correlations experienced across exposure stimuli impacts the perceptual weight of F0 in /b/–/p/ categorization. We examined the proportion of *pier* responses to test stimuli across a 2 (F0) × 3 (Block) × 2 (Talker) ANOVA. This revealed a robust main effect of F0, $F(1, 24) = 64.06$, $p < .001$, $\eta_p^2 = .73$, and no effects of either block, $F(2, 48) = 2.82$, $p = .069$, $\eta_p^2 = .11$, or talker ($F < 1$). There was a robust interaction between F0 and block, $F(2, 48) = 26.28$, $p < .001$, $\eta_p^2 = .52$, indicating that the short-term F0 × VOT regularities impacted the effectiveness of F0 on /b/–/p/ categorization. There was also an interaction of F0 and talker, $F(1, 24) = 62.29$, $p < .001$, $\eta_p^2 = .72$, indicating a difference in reliance on F0 for /b/–/p/ categorization across the implied talkers. Neither the Block × Talker interaction ($F < 1$) nor the three-way interaction, $F(2, 48) = 1.22$, $p = .303$, $\eta_p^2 = .05$, was significant.

We next examined the simple interaction between F0 and block individually for each talker (high/low F0 range, female/male video) to determine whether the modulation of the influence of F0 on /b/–/p/ categorization occurred across each of the "talkers" implied by the two videos (Figure 9a, 9b). For the acoustic stimuli sampling the high F0 range paired with the female video, there were significant main effects of F0, $F(1, 24) = 105.49$, $p < .001$, $\eta_p^2 = .82$, and block, $F(1, 24) = 3.89$, $p = .027$, $\eta_p^2 = .14$. As is evident in Figure 9a, there was a robust interaction between F0 and block, $F(2, 48) = 11.18$, $p < .001$, $\eta_p^2 = .32$. F0 exerted a significant influence on test stimulus categorization in each of the blocks: female$_{CAN}$–male$_{CAN}$ block, $F(1, 24) = 129.20$, $p < .001$; female$_{REV}$–male$_{CAN}$, $F(1, 24) = 31.26$, $p < .001$; female$_{CAN}$–male$_{REV}$, $F(1, 274) = 58.12$, $p < .001$. Importantly, the magnitude of the influence of F0 was reduced upon introduction of the artificial accent (reverse F0 × VOT correlation) to the female speaker implied by stimulus sampling and video. The influence of F0 effect was significantly less robust in the female$_{REV}$–male$_{CAN}$ block (M = .26, SE = .05, 95% CI [.17, .36]) than in the female$_{CAN}$–male$_{CAN}$ block (M = .51, SE = .05, 95% CI [.42, .60], $p < .001$) or female$_{CAN}$–male$_{REV}$ block (M = .41, SE = .05, 95% CI [.30, .52], $p = .027$). Listeners relied less on F0 in categorizing the test stimuli upon introduction of the artificial accent in the female$_{REV}$–male$_{CAN}$ block, consistent with dimension-based statistical learning across the implied female talker even in the context of opposing statistics within the block from the male talker.

Figure 9b shows the results for the low F0 range stimuli paired with the male video. The influence of F0 was marginal, $F(1, 24) = 3.35$, $p = .080$, $\eta_p^2 = .12$. The main effect of block was not significant ($F < 1$). However, the F0 × Block interaction was significant, $F(2, 48) = 17.83$, $p < .001$, $\eta_p^2 = .43$, indicating that the short-term regularities present across blocks influenced the effectiveness of F0 in /b/–/p/ categorization. Specifically, the influence of F0 was significant in both the female$_{CAN}$–male$_{CAN}$, $F(1, 24) = 13.50$, $p = .001$, and female$_{CAN}$–male$_{REV}$, $F(1, 24) = 22.75$, $p < .001$ blocks, although in opposing directions. As in Experiment 2, the influence of F0 mirrored the English regularity in the female$_{CAN}$–male$_{CAN}$ block, and was reversed in the female$_{CAN}$–male$_{REV}$ block. In the context of an acoustic artificial accent paired with a male video, listeners' reliance on F0 mirrored the short-term regularity of the artificial accent. When the F0/VOT correlation was reversed for the male talker, F0 was used in a way that mirrored the local acoustical statistics, a pattern that was not seen for the female talker. The F0 effect was not significant in the female$_{REV}$–male$_{CAN}$ block, $F(1, 24) = 5.52$, $p = .027 > .013$ (with alpha adjusted for three comparisons), but the effect size of F0 did not differ significantly between the female$_{CAN}$–male$_{CAN}$ block (M = .17, SE = .05, 95% CI [.07, .26]) and female$_{REV}$–male$_{CAN}$ block (M = .13, SE = .06, 95% CI [.02, .25]) blocks ($p > .999$), each of which was significantly different from the female$_{CAN}$–male$_{REV}$ block (M = −.12, SE = .03, 95% CI [–.18, −.07]; ps < .001).

Finally, we examined whether participants responded differently to the test stimuli with the same absolute F0 value (210 Hz) and identical voice acoustics as a function of the implied talker. As shown in Figure 9c, a 3 (Block) × 2 (Talker) ANOVA for the middle-range F0 test stimulus (210 Hz F0) across trials for which it was paired with the female versus the male video revealed a significant main effect of block, $F(2, 48) = 5.48$, $p = .007$, $\eta_p^2 = .19$, and talker, $F(1, 24) = 1.77$, $p < .001$, $\eta_p^2 = .68$, and a significant interaction, $F(2, 48) = 9.01$, $p < .001$, $\eta_p^2 = .27$. The simple effect analyses indicate that acoustically identical test stimuli are categorized more often as *pier* when paired with the male video than when the same stimulus is paired with the female video across experimental blocks, ps < .010. Figure 9c summarizes the influence of F0 across the implied talkers.

In all, the results of Experiment 3 demonstrate that dimension-based statistical learning is supported by visual information conveying talker identity even when the acoustic dimensions conveying an artificial accent are identical across the implied "talkers." The distinct face information provided in the videos was sufficient to allow listeners to track the distinct acoustic regularities con-

[1] The careful reader will note the odd data point for the high female video at 20 ms VOT. We returned to the raw data to investigate. It is unusual, but verifiable.

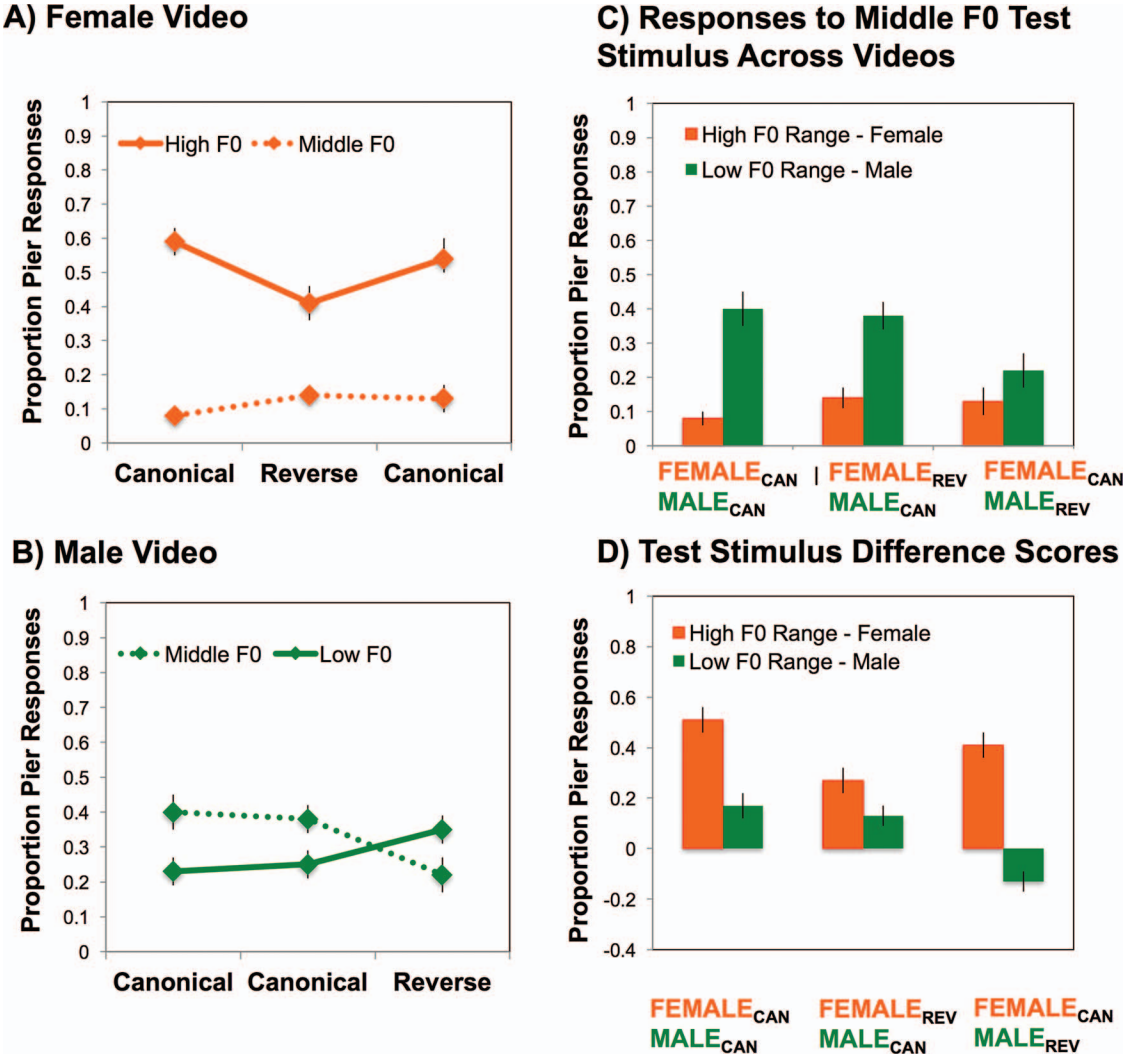

Figure 9. Results of Experiment 3. Mean proportion *pier* responses to Experiment 3 test stimuli as a function of the sampling of the range of fundamental frequency (F0; high, low) and the video accompanying the acoustic speech stimuli (female, male) and the short-term regularity of the exposure stimuli (canonical [CAN], reverse [REV]). (A) Proportion of *pier* responses to the high (290 Hz) and middle (210 Hz) F0 test stimuli as a function of the short-term regularity experienced paired with the female video. (B) Proportion of *pier* responses to the middle (210 Hz) and low (130 Hz) F0 test stimuli as a function of the short-term regularity experienced in the low F0 range and paired with the male video. Note that the dashed lines in Panels (A) and (B) correspond to responses to the middle F0 test stimulus with the same F0 and voice onset time characteristics. (C) Mean proportion of *pier* responses to the middle F0 (210 Hz) test stimulus for the female and male voices. (D) Summary of the contribution of F0 to *beer–pier* categorization responses across conditions, plotted as the difference score between the two test stimuli in each condition. See the online article for the color version of this figure.

veyed across the high F0 and low F0 ranges. Rather than a global accumulation of acoustic dimensional regularities, the system tracks simultaneously evolving short-term distributional statistics when there is information available to indicate how to "bin" the incoming instances. This is true even in the case when there is a single acoustic voice conveying the conflicting distributional regularities.

General Discussion

Prior research has demonstrated that short-term distributional regularities strongly impact the effectiveness of acoustic dimensions in signaling phonetic categories (Idemaru & Holt, 2011, 2014; Lehet & Holt, 2017; Liu & Holt, 2015). For example, in the context of an artificial accent that reverses the typical English relationship of F0 and VOT, such that higher F0s are paired with shorter VOTs and lower F0s are paired with longer VOTs, listeners rapidly down-weight reliance upon F0. This highlights that the influence of acoustic dimensions on phonetic categorization is not fixed; rather, it adapts flexibly in response to regularities in short-term speech input. However, the information listeners use to track regularities across specific talkers, or groups of talkers is not

well-understood in dimension-based statistical learning, or in adaptive plasticity in speech perception more generally.

One advantage of investigating adaptive plasticity in speech perception using a dimension-based statistical learning approach is that it provides a direct measure of online recalibration of the effectiveness of specific acoustic dimensions in signaling speech categories. With it, it is possible to measure adaptive changes in the contribution of specific acoustic dimensions as a function of the regularities of short-term speech input. Another advantage is that it is possible to manipulate detailed regularities across these dimensions in the input. Here, we capitalized on these features to investigate the information that contributes to tracking simultaneously evolving regularities in the acoustic dimensions that signal speech categories across time. The case under investigation here was a rather strict test in that it involved a complete reversal of the covariation typically present across two acoustic dimensions signaling English /b/ and /p/ categories, F0 and VOT.

In Experiment 1, we observed that the relative range of an acoustic dimension experienced affects how the dimension contributes to speech categorization. Said another way, what counts as a "low F0" differs as a function of the range of experienced F0 frequencies. In Experiment 1, an intermediate F0 frequency experienced across a block of trials sampling a lower range of F0 more often signaled /p/ than when the same stimulus was experienced across a block of trials sampling a higher F0 range. Critically, this had implications for adaptive plasticity. The pattern of dimension-based statistical learning observed across the groups was affected by the *relative,* rather than absolute, values experienced across the F0 dimension.

Although a primary goal of Experiment 1 was to establish a testing ground for Experiments 2 and 3, this finding is significant in its own right. The range of variability experienced across an acoustic dimension, here F0, influenced the manner in which it contributed to speech categorization. In the context of higher F0 exemplars, middle F0 tokens were perceived more often as /b/, consistent with the long-term mapping of lower F0s to /b/ compared to /p/. The same middle F0 tokens were more often perceived as /p/ in the context of lower F0 exemplars. This is a form of distributional learning that affects both online speech categorization and also adaptive plasticity, as shown in Figures 2 and 3, respectively. These results emphasize the need for dynamic models of speech categorization; the mapping of acoustic input dimensions to phonetic categories is not static. Moreover, they highlight the interconnected nature of perceptual effects the literature tends to dissociate as normalization effects (e.g., Figure 2) versus adaptive plasticity (e.g., Figure 3), a matter discussed recently by Weatherholtz and Jaeger (2016). Speech categorization has long been known to be context-dependent (Ladefoged & Broadbent, 1957; Lindblom & Studdert-Kennedy, 1967), with diverse sources of context information affecting speech categorization (e.g., Bertelson et al., 2003; Dilley & Pitt, 2010; Ganong, 1980; Holt, 2005; Lotto & Kluender, 1998; Miller & Liberman, 1979). Experiment 1 establishes that the distributional sampling of acoustic dimensions across a stimulus set is sufficient to shift speech category boundaries. Even when listeners categorize isolated speech stimulus tokens that are otherwise "context free," the set of stimuli within which the tokens are presented serves as a statistical context that influences categorization. Finally, Experiment 1 extended prior research to demonstrate that dimension-based statistical learning across the F0 × VOT acoustic space is evident in a lower frequency F0 range. As in the higher frequency range investigated here and in prior work (Idemaru & Holt, 2011, 2014; Lehet & Holt, 2017), the introduction of the artificial accent with an F0 × VOT correlation opposite that typical of English led to reduced reliance on F0 in /b/–/p/ categorization across stimuli sampling a lower F0 range; this provides an important case for comparison with the later experiments.

Dimension-based statistical learning provides a window through which to view the dynamic remapping of how acoustic information informs phonetic categorization. In the sense that it involves the mapping of relatively early acoustic dimensions to phonetic categories, it is tempting to imagine that this form of adaptive plasticity in speech perception might simply involve the accumulation of low-level acoustic distributional regularities. If it were, the competing F0 × VOT correlations (canonical, reverse) presented simultaneously in Experiments 2 and 3 would cancel in the global acoustic distribution. From the perspective of distributions of experience across the F0 and VOT acoustic dimensions there was no F0 × VOT correlation in short-term speech input. Because these global statistics do not change across blocks, perception that is driven solely by accumulation of acoustic instances across the F0 × VOT acoustic space should not by modulated across blocks.

Instead, the opposing local distributional regularities that jointly defined the global acoustic regularity impacted speech categorization. Experiments 2 and 3 demonstrated robust down-weighting of F0 that patterned with the introduction of the artificial accent (reverse). This was supported by the availability of information about the sampling range of the F0 dimension (as observed in Experiment 1), the presence of distinct talkers conveyed through voice quality (Experiment 2), and the presence of visual information consistent with different talkers (Experiment 3). Even when the acoustic speech signals were drawn from the same voice (Experiment 3), it was possible for listeners to track simultaneous short-term regularities in the F0 × VOT acoustic space when the distinct distributions were paired with silent videos depicting different talkers. In this case, the patterns of down-weighting of the F0 dimension observed could only arise from learning about the distributions of speech sounds associated with the visually cued attributes of the two talkers presented in the videos. These conditions model some real demands of listening environments, within which it is common to experience speech from multiple talkers interleaved such that their distinct short-term regularities evolve together across time. For example, studies find effects of accent strength and familiarity with an accent on the degree to which speech perception adapts to foreign-accented speech (Witteman, Weber, & McQueen, 2013). Indirectly, this suggests that listeners must have some ability to track multiple regularities. The present results directly demonstrate listeners' ability to track coevolving regularities—even when short-term regularities are actually *opposing* one another. This provides a rather stringent test of listeners' ability to track the simultaneously evolving short-term regularities.

It is tempting to consider these results as evidence of speaker-specific adaptive plasticity, because the information sources available to support the binning of local statistical distributions were voice quality in Experiment 2 and talker identity conveyed via articulating faces in Experiment 3. Both the voice quality and the videos provided information that a specific speech tokens experi-

enced across the F0 × VOT acoustic space contribute to a specific sampling of the statistical regularities for that "talker." Indeed, research investigating adaptive plasticity in speech perception using other experimental approaches has documented cases of talker-specific phonetic tuning. For example, the lexically driven perceptual tuning of ambiguous fricatives driven by experience with one talker does not always generalize to other talkers (Eisner & McQueen, 2005; Kraljic & Samuel, 2005, 2006, 2007).

Nonetheless, it is important to acknowledge the very real possibility that the origin of the influence we observe in the present studies may be less about talker and more about simply having a consistent external information source to support the organization of acoustic regularities. From this perspective, one might ask whether *any* consistent information source might serve similarly and, in this regard, the results of Idemaru and Holt (2014) are revealing. Manipulating F0 × VOT regularities across blocks as in the present experiments, Idemaru and Holt found that listeners track acoustic F0 × VOT regularities independently across place-of-articulation. When an artificial accent reversing the correlation between F0 and VOT was introduced for /b/–/p/ stimuli in a block of trials that also included a canonical F0 × VOT correlation for /d/–/t/ stimuli, listeners down-weighted reliance on F0 for /b/–/p/ while maintaining reliance on F0 in /d/–/t/ categorization for the *same voice*. Idemaru and Holt speculated that the ability to track opposing statistics across /b/–/p/ and /d/–/t/ categories could arise from the differential sampling of the range of VOT values across place-of-articulation (similar to the differential ranges of F0 in the present experiments), or acoustic information in the signals that reveals place-of-articulation such as the spectral shape of the initial consonant burst. Thus, like the present studies, an additional information source appears to have provided evidence allowing the system to effectively bin overlapping acoustic regularities. Though the present studies used information closely tied to talker as it models important information available to listeners in complex speech environments, talker information should not be viewed as the only information source available to help listeners parse simultaneously coevolving distributional regularities.

Indeed, even in lexically driven perceptual tuning paradigms that have observed talker-specificity in adaptive plasticity, the precise sampling of speech tokens across perceptual space has a large influence on whether generalization is observed. Generalization across voices can be encouraged or discouraged by simply sampling different ranges of the specific voices' perceptual space, with generalization observed only when voices sample a common perceptual space (Reinisch & Holt, 2014). This may at least partly explain why even when generalization is observed it is often less robust than that observed for the talker experienced in exposure (Liu & Holt, 2015; van der Zande, Jesse, & Cutler, 2013) The present results demonstrate that information to support the binning of acoustic regularities allows listeners to track multiple regularities in short-term speech input, even when those regularities present opposing statistics. Thus, the present results demonstrate that tracking distributional regularities across acoustic dimensions, and perceptual adjustment of the weight of those dimensions in perceptual categorization is one means by which listeners' sensitivity to talker-specific patterns of speech may influence phonetic categorization. The present results indicate that rather than a simple accumulation of instances across low-level perceptual dimensions, this adaptive plasticity depends upon additional information available in the acoustic and visual input to provide the context with which to bin across competing short-term regularities. This is in accord with studies of lexically driven adaptive plasticity that listeners' perception of the situation, for example whether there is one talker or two, plays a role above and beyond the acoustics of the speech alone (Samuel & Kraljic, 2009; Kraljic & Samuel, 2007, 2005).

This point is useful to consider in light of the patterns of adaptive plasticity observed for the "male" talker suggested by the input in Experiments 2 and 3, especially in the context of the results of Experiment 1. The results of Experiments 2 and 3 demonstrate that distinct distributions can be tracked when there is information available to support the binning of the regularities and that these evolving distributions impact how acoustic dimensions contribute to speech categorization. But, are the distributions tracked *independently?* The pattern of results in Experiment 1, in comparison to Experiments 2 and 3, suggests that there is some cross-talk. In Experiment 1, different groups of participants experienced distributions sampling the high range of F0 and the low range of F0. In contrast, participants in Experiments 2 and 3 experienced both samples, intermixed across trials within a block and paired with information to support differential binning of the distributions. In fact, Experiment 3 involved the same acoustic information as Experiment 1, albeit mixed within presentation instead of segregated across participants. The acoustic distributions sampling the higher range of F0 and paired with female talker information from voice (Experiment 2) or face (Experiment 3) led to consistent effects of F0 on /b/–/p/ categorization and clear evidence of down-weighting consistent with that observed in Experiment 1. Even on its own, this demonstrates listeners' ability to effectively bin evolving short-term speech regularities; had they been sensitive only to the sampling of instances across F0 and VOT, the presence of opposing statistics associated with the male talker would have eliminated the F0 × VOT correlation and the adaptive plasticity that tracks with it.

Yet, in contrast to the female talker, stimuli sampling the lower F0 range paired with information consistent with a male talker consistently showed a weaker F0 effect when male F0 × VOT correlations were canonical. In addition, when the relationship between F0 and VOT was reversed in short-term speech input, the F0 effect flipped to reflect the short-term regularity. Prior studies (all conducted with a female voice quality and a higher F0 range, with no concurrent talker; Idemaru & Holt, 2011, 2014) had not observed a reversal of the influence of F0 on /b/–/p/ categorization to mirror the short-term input statistics even after 5 days of experience with the reversed F0 × VOT correlation (Idemaru & Holt, 2011). Most intriguing, in the present studies, this pattern held for the male talker only when it was mixed in presentation with the higher F0 range speech tokens (Experiments 2 and 3), and not in Experiment 1 for which participants experienced only the lower range of F0 without additional information to imply a male talker (either by voice quality or gender of the talker in a video). One explanation may be that the voice quality of the original, Experiment 1, speech tokens were more consistent with a female voice because the original talker upon whom the stimuli were based was female. Although speculative, it is also intriguing to consider this pattern in light of recent observations of the

F0 × VOT correlations present in native-English talkers' speech productions. In a study on dimension-based statistical learning on speech *production*, Lehet and Holt (2017) observed that male participants failed to use F0 to reliably differentiate /b/–/p/ categories in their own speech productions whereas female participants consistently used F0 systematically in the canonical manner across /b/–/p/ categories. Caution is warranted as this observation arises from acoustic measurements across a small sample of talkers. Yet, if it proves to be hold more widely that F0 with VOT covary less reliably in male English talkers' /b/–/p/ productions, then it is possible that native-English listeners have developed different expectations about the robustness of F0 × VOT correlations across male versus female talkers that affect adaptive plasticity. In line with conclusions from Kraljic and Samuel (2007), this would suggest that when acoustics are ambiguous (as in Experiment 3) listeners rely more on speaker-specific learning. Our results suggest the possibility that expectations of about different groups of listeners' speech productions might be a source of influence.

The present findings also indicate that refinement is needed in the way that we think about acoustic dimensions as contributing to phonetic categorization. It is easy to speak of the adaptive plasticity observed in experiencing the reversal of the F0 × VOT correlation in short-term input as "down-weighting of F0." Prior research has demonstrated that we cannot equate this with a wholesale turn of attention away from the F0 dimension because listeners quickly adapt to again rely on F0 for phonetic categorization when the short-term regularity shifts back to a canonical correlation (Idemaru & Holt, 2011); therefore, they must continue to track F0 even as it is down-weighted in its influence on phonetic categorization. The present results suggest that it may be possible that dynamic reweighting is happening not across an entire dimension (e.g., F0), but rather across a *specific range* of an acoustic dimension. Future research that manipulates the experienced range of values along a dimension and the generalization of down-weighting to speech exemplars sampling the dimension outside the experienced range will be informative to refining this issue.

Input regularities experienced in the local environment are sometimes at odds with the distributional regularities of long-term experienced that establish category representations. This situation is illustrated quite clearly in encountering accented speech, but it is a more general challenge for all perceptual systems. The present results highlight that the dynamic reweighting of perceptual dimensions' influence on categorization in response to short-term input regularities is not a result of a simple accumulation of sensory instances. Rather, perceivers are able to track multiple, context-sensitive regularities simultaneously, with rapid context-dependent adjustments that impact the relative effectiveness of perceptual dimensions in signaling categories.

References

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

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

Boersma, P., & Weenink, D. (2017). *Praat: Doing phonetics by computer* (Version 6.0.29) [Computer program]. 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. http://dx.doi.org/10.1006/jpho.1996.0021

Clarke-Davidson, C. M., Luce, P. A., & Sawusch, J. R. (2008). Does perceptual learning in speech reflect changes in phonetic category representation or decision bias? *Perception & Psychophysics, 70,* 604–618. http://dx.doi.org/10.3758/PP.70.4.604

Dilley, L. C., & Pitt, M. A. (2010). Altering context speech rate can cause words to appear or disappear. *Psychological Science, 21,* 1664–1670. http://dx.doi.org/10.1177/0956797610384743

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

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41,* 1149–1160. http://dx.doi.org/10.3758/BRM.41.4.1149

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. http://dx.doi.org/10.1121/1.2945161

Ganong, W. F., III. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance, 6,* 110–125. http://dx.doi.org/10.1037/0096-1523.6.1.110

Guediche, S., Blumstein, S. E., Fiez, J. A., & Holt, L. L. (2014). Speech perception under adverse conditions: Insights from behavioral, computational, and neuroscience research. *Frontiers in Systems Neuroscience, 7,* 126. http://dx.doi.org/10.3389/fnsys.2013.00126

Holt, L. L. (2005). Temporally nonadjacent nonlinguistic sounds affect speech categorization. *Psychological Science, 16,* 305–312. http://dx.doi.org/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. http://dx.doi.org/10.1121/1.2188377

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

Idemaru, K., & Holt, L. L. (2014). Specificity of dimension-based statistical learning in word recognition. *Journal of Experimental Psychology: Human Perception and Performance, 40,* 1009–1021. http://dx.doi.org/10.1037/a0035269

Kim, M., & 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. http://dx.doi.org/10.1353/lan.1994.0023

Kraljic, T., Brennan, S. E., & Samuel, A. G. (2008). Accommodating variation: Dialects, idiolects, and speech processing. *Cognition, 107,* 54–81. http://dx.doi.org/10.1016/j.cognition.2007.07.013

Kraljic, T., & Samuel, A. G. (2005). Perceptual learning for speech: Is there a return to normal? *Cognitive Psychology, 51,* 141–178. http://dx.doi.org/10.1016/j.cogpsych.2005.05.001

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

Kraljic, T., & Samuel, A. G. (2007). Perceptual adjustments to multiple speakers. *Journal of Memory and Language, 56,* 1–15.

Kraljic, T., Samuel, A. G., & Brennan, S. E. (2008). First impressions and last resorts: How listeners adjust to speaker variability. *Psychological Science, 19,* 332–338. http://dx.doi.org/10.1111/j.1467-9280.2008.02090.x

Ladefoged, P., & Broadbent, D. E. (1957). Information conveyed by vowels. *The Journal of the Acoustical Society of America, 29,* 98–104. http://dx.doi.org/10.1121/1.1908694

Lehet, M., & Holt, L. L. (2017). Dimension-based statistical learning affects both speech perception and production. *Cognitive Science, 41,* 885–912. http://dx.doi.org/10.1111/cogs.12413

Lindblom, B. E., & Studdert-Kennedy, M. (1967). On the role of formant transitions in vowel recognition. *The Journal of the Acoustical Society of America, 42,* 830–843. http://dx.doi.org/10.1121/1.1910655

Liu, R., & Holt, L. L. (2015). Dimension-based statistical learning of vowels. *Journal of Experimental Psychology: Human Perception and Performance, 41,* 1783–1798. http://dx.doi.org/10.1037/xhp0000092

Lotto, A. J., & Kluender, K. R. (1998). General contrast effects in speech perception: Effect of preceding liquid on stop consonant identification. *Perception & Psychophysics, 60,* 602–619. http://dx.doi.org/10.3758/BF03206049

McMurray, B., Tanenhaus, M. K., & Aslin, R. N. (2002). Gradient effects of within-category phonetic variation on lexical access. *Cognition, 86,* B33–B42. http://dx.doi.org/10.1016/S0010-0277(02)00157-9

Miller, J. L., & Liberman, A. M. (1979). Some effects of later-occurring information on the perception of stop consonant and semivowel. *Perception & Psychophysics, 25,* 457–465. http://dx.doi.org/10.3758/BF03213823

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

Reinisch, E., & Holt, L. L. (2014). Lexically guided phonetic retuning of foreign-accented speech and its generalization. *Journal of Experimental Psychology: Human Perception and Performance, 40,* 539–555. http://dx.doi.org/10.1037/a0034409

Samuel, A. G., & Kraljic, T. (2009). Perceptual learning for speech. *Attention, Perception & Psychophysics, 71,* 1207–1218. http://dx.doi.org/10.3758/APP.71.6.1207

Schertz, J., Cho, T., Lotto, A., & Warner, N. (2016). Individual differences in perceptual adaptability of foreign sound categories. *Attention, Perception & Psychophysics, 78,* 355–367. http://dx.doi.org/10.3758/s13414-015-0987-1

Trude, A., & Brown-Schmidt, S. (2012). Talker-specific perceptual adaptation during online speech perception. *Language and Cognitive Processes, 27,* 979–1001. http://dx.doi.org/10.1080/01690965.2011.597153

van der Zande, P., Jesse, A., & Cutler, A. (2013). Lexically guided retuning of visual phonetic categories. *The Journal of the Acoustical Society of America, 134,* 562–571. http://dx.doi.org/10.1121/1.4807814

Vroomen, J., & Baart, M. (2009). Phonetic recalibration only occurs in speech mode. *Cognition, 110,* 254–259. http://dx.doi.org/10.1016/j.cognition.2008.10.015

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

Weatherholtz, K., & Jaeger, T. F. (2016). Speech perception and generalization across speakers and accents. In M. Aronoff (Ed.), *Oxford Research Encyclopedia of Linguistics*. http://dx.doi.org/10.1093/acrefore/9780199384655.013.95

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

Witteman, M. J., Weber, A., McQueen, J. M. (2013). Foreign accent strength and listener familiarity with an accent codetermine speed of perceptual adaptation. *Attention, Perception, and Psychphysics, 75,* 53.

Received March 15, 2018
Revision received May 9, 2018
Accepted May 16, 2018 ■