Dimension-based statistical learning affects both speech perception and production

Matthew Lehet

and

Lori L. Holt

Carnegie Mellon University

Department of Psychology and the Center for Neural Basis of Cognition

Author Note

This research was supported by a grant to LLH from the National Institutes of Health (R01DC004674) and institutional training grants supporting ML (T90DA022761, T32GM081760). The authors thank Dr. Kaori Idemaru, Monica Li and Christi Gomez for their contributions to the research.

Correspondence should be addressed to Matthew Lehet, Department of Psychology, Carnegie Mellon University, Pittsburgh PA, 15213, mil@andrew.cmu.edu

Abstract

Multiple acoustic dimensions signal speech categories. However, dimensions vary in their informativeness; some are more diagnostic of category membership than others. Speech categorization reflects these dimensional regularities such that diagnostic dimensions carry more "perceptual weight" and more effectively signal category membership to native listeners. Yet, perceptual weights are malleable. When short-term experience deviates from long-term language norms, such as in a foreign accent, the perceptual weight of acoustic dimensions in signaling speech category membership rapidly adjusts. The present study investigated whether rapid adjustments in listeners' *perceptual* weights in response to speech that deviates from the norms also affects listeners' own speech *productions*. In a word recognition task, the correlation between two acoustic dimensions signaling consonant categories, fundamental frequency (F0) and voice onset time (VOT), matched the correlation typical of English, then shifted to an "artificial accent" that reversed the relationship, and then shifted back. Brief, incidental exposure to the artificial accent caused participants to down-weight perceptual reliance on F0, consistent with previous research. Throughout the task, participants were intermittently prompted with pictures to produce these same words. In the block in which listeners heard the artificial accent with a reversed F0 x VOT correlation, F0 was a less robust cue to voicing in listeners' own speech productions. The statistical regularities of short-term speech input affect both speech perception and production, as evidenced via shifts in how acoustic dimensions are weighted.

Keywords: Speech recognition, Perception, Communication, Language Understanding, Motor control, Perceptual Weighting, Dimension-based Statistical Learning

Ambient speech affects speech production. The acoustic details of the speech we produce are affected by vocal feedback from our own voice, and also from speech overheard from other talkers. Adjustments to speech production in response to auditory input from one's own voice are very clear in sensorimotor adaptation paradigms. When speech is artificially distorted and fed back with a negligible delay through headphones, talkers quickly adapt by producing speech with acoustics adjusted in a direction that opposes the distortion. Experiencing vocal feedback with an artificially increased fundamental frequency (F0), for example, leads talkers to *decrease* F0 in subsequent speech productions (Donath, Natke, & Kalveram, 2002; Houde & Jordan, 1998; Nasir & Ostry, 2009; Purcell & Munhall, 2006; Villacorta, Perkell, & Guenther, 2007). Intriguingly, recent research demonstrates that the consequences of sensorimotor adaptation extend to perception of other talkers' speech. Following sensorimotor adaptation, perceptual category boundaries for other talker's speech are shifted relative to baseline perceptual boundaries measured prior to adaptation. The direction of the shift mirrors the direction of adaptation (Lametti, Rochet-Capellan, Neufeld, Shiller, & Ostry, 2014b; Shiller, Sato, Gracco, & Baum, 2009), and the effects appear to be reciprocal. Perceptual shifts evoked by carrier phrases or explicit feedback-based perceptual training on another talker's speech can also impact the degree of sensorimotor adaptation to distorted vocal feedback from one's own voice (Bourguignon, Baum, & Shiller, 2016; Lametti, Krol, Shiller, & Ostry, 2014a; Shiller, Lametti, & Ostry, 2013).

Speech production is also influenced by speech input from other talkers. For example, when participants rapidly repeat (shadow) recorded speech from another talker (Goldinger, 1998; Marslen-Wilson, 1973), they tend to come to imitate acoustic details of the shadowed talker (Fowler, Brown, Sabadini, & Welhing, 2003; Goldinger, 1998; Honorof, Weihing, & Fowler,

2011; Miller, Sanchez, & Rosenblum, 2013; Roon & Gafos, 2014; Shockley, Sabadini, & Fowler, 2004). Speech experienced in more natural conversational interactions can also influence speech production (Coupland & Giles, 1988; Giles, 1973; 1977; Pardo, 2006; Pardo, Gibbons, Suppes, & Krauss, 2012; Pardo, Jay, & Krauss, 2010). Likewise, experience in second language learning environments appears to leave its mark on language-learners' first-language speech productions (Chang, 2012; 2013; Guion, 2002; Lord, 2008; Sancier & Fowler, 1997), and exposure to speech that departs from the norms of the language community, such as encounters with foreign-accented speech, can result in changes to speech production that mimic the experienced accent (Delvaux & Soquet, 2007).

This brief sampling of the literature illustrates the breadth of evidence that heard speech affects speech production. Listeners appear to track detailed acoustic information from speech input such that it has an impact on how acoustic dimensions are realized in subsequent speech productions. But, many open questions remain. In the present study, we examine *cue weighting* as a phenomenon that may provide empirical leverage in seeking to better understand how other talkers' speech can impact one's own speech productions.

The concept of cue weighting in speech perception and production arises in examining the acoustic dimensions that signal speech categories in a language community and how listeners make use of them in speech perception (e.g., Holt & Lotto, 2006). Speech acoustics are notoriously multidimensional. However, the contributing dimensions are not necessarily equally diagnostic of phonetic category membership (Francis, Baldwin, & Nusbaum, 2000; Francis, Kaganovich, & Driscoll-Huber, 2008; Holt & Lotto, 2006; Idemaru, Holt, & Seltman, 2012; Iverson & Kuhl, 1995; Nittrouer, 2004; Shultz, Francis, & Llanos, 2012). For example, in American English both the second and the third formant onset frequencies vary across /r/ and /l/.

However, the third formant is much more robustly associated with /r/-/l/ category membership than the second formant (Idemaru & Holt, 2013; Lotto, Sato, & Diehl, 2004). By adulthood, native listeners' speech perception reflects these subtle dimensional regularities of native speech productions. In categorizing /r/-/l/ sounds varying in both the second and third formant onset frequencies, native English listeners rely much more on the highly-diagnostic third formant frequency, giving it greater *perceptual weight* than the second formant frequency (Ingvalson, McClelland, & Holt, 2011; Iverson et al., 2003; Yamada & Tohkura, 1992). Perceptual weights appear to be built up over a long developmental course extending into late childhood or early adolescence (Hazan & Barrett, 2000; Idemaru & Holt, 2013; Lowenstein & Nittrouer, 2008; Nittrouer, 2004; Nittrouer, Lowenstein, & Packer, 2009) and, once established, they are quite stable (Idemaru et al., 2012). Thus, at least within a language community, there is a rather close correspondence between the long-term regularities of how acoustic dimensions relate to phonetic categories across speech productions and the weight of listeners' reliance on these dimensions in speech perception.

Yet, listeners often encounter foreign accents, dialects, or significant background noise. These factors 'warp' speech acoustics relative to the long-term language community norms. Under these conditions perceptual weights that reflect long-term language regularities are suboptimal in the short-term. However, it appears that although perceptual weights are quite stable when measured under conditions that mimic long-term regularity, they are also malleable in response to short-term input that deviates from long-term language norms. Perceptual weights rapidly adjust in online speech perception. For example, Idemaru and Holt (Idemaru & Holt, 2011; 2014; see also Liu & Holt, 2015) introduced an "artificial accent" to spoken words by manipulating the correlation between two acoustic dimensions significant in signaling the

consonant voicing differences between the rhymes *beer/pier* and *deer*/*tear.* One of these dimensions, (voice onset time (VOT), the duration between acoustic evidence of consonant release and acoustic evidence of vocal cord vibration) is given greater perceptual weight compared to the other dimension (fundamental frequency of the following vowel, F0), although each contributes to categorization of the consonant as /b/ versus /p/ or /d/ versus /t/ (Francis et al., 2008; Lisker, 1986; Lisker & Abramson, 1985; Whalen, Abramson, Lisker, & Mody, 1993). In English and many other languages (Diehl & Kingston, 1991), voiced consonants (like *beer* and *deer*) tend to have shorter voice onset times (VOTs) and tend to precede vowels with lower F0s, whereas voiceless consonants (like *pier* and *tear*) with longer VOTs are associated with vowels realized with higher F0s (Castleman & Diehl, 1996; Lisker & Abramson, 1985). When both VOT and F0 vary across speech productions, native English listeners rely primarily on the VOT dimension to signal category membership, but F0 plays a secondary role (Francis et al., 2008; Lisker & Abramson, 1985). This is very evident when VOT is perceptually ambiguous. In this case, a higher F0 robustly signals voiceless consonants (*pier, tear*) whereas a lower F0 signals voiced consonants (*beer, deer*), consistent with the long-term regularities of English speech productions (Kingston & Diehl, 1994; Kohler, 1982; 1984).

Throughout the Idemaru and Holt (2011) experiment, listeners simply identified each spoken word as rhymes *beer, deer, pier,* or *tear*. For the majority of trials, the word identities were signaled unambiguously by VOT. For these Exposure trials, the relationship between VOT and F0 varied subtly across blocks. Some blocks included stimuli that followed the canonical English F0 x VOT relationship (higher F0s for longer VOTs, lower F0s for shorter VOTs; (Kingston & Diehl, 1994). Participants also heard a block of trials with an "artificial accent" that reversed this relationship (lower F0s with longer VOTs, higher F0s with shorter VOTs).

Throughout the experiment, participants also heard a small proportion of Test stimuli with a perceptually ambiguous VOT; the Test stimuli could be disambiguated only by F0. Idemaru and Holt found that F0 robustly influences Test-trial categorization in Canonical blocks in which the majority of stimuli match long-term English experience with F0 and VOT. This reflects listeners' long-term experience with F0 as a secondary, but significant, cue in signaling these consonants in English input. However, upon introduction of the artificial accent that reversed the F0 x VOT relationship, F0 rapidly became less effective at signaling category membership for Test trials; its perceptual weight decreased.

Listeners were not informed about the artificial accent, the voice remained constant, the blocks were not differentiated in any way to participants, and the task was always simply to identify the word. The range of dimension variability experienced across blocks fell within that experienced for individual talkers, and it went largely unnoticed by participants. In these ways, the learning that underlies the rapid adjustment of perceptual weights is incidental.

These results indicate that listeners track the relationship of acoustic dimensions across short-term input and that sensitivity to short-term regularities across dimensions affects perceptual weights. Listeners are highly sensitive to evolving dimensional regularities in the short-term input and short-term deviations of regularities from the norm. Deviations, such as in encountering an accent, result in rapid adjustments in the perceptual weight, or influence, of acoustic dimensions on speech categorization. Idemaru and Holt (2011) refer to this incidental learning as *dimension-based statistical learning*.

There is a close relationship between the regularities experienced across native speech productions and the perceptual weights that native listeners apply in perception. Moreover, other research demonstrates that the details of other talkers' speech can influence one's own speech

productions (Coupland & Giles, 1988; Delvaux & Soquet, 2007; Pardo, 2013). However, it is not yet known if the rapid adjustments of *perceptual* cue weights that arise from short-term deviations in speech input have any effect on listeners' own speech *productions*. Understanding whether rapid adjustments in perceptual weights have concomitant consequences on speech production presents the opportunity to examine fine-grained interactions of speech perception and production using perceptual weights as a tool.

In the present study, we incorporate a speech production task into the perceptual paradigm of Idemaru and Holt (2011) to investigate perception-production interactions in a context in which learning is driven by acoustic cue relationships akin to those that might be encountered in incidental listening to accented speech. The paradigm provides the opportunity to directly manipulate acoustic dimensions in a manner unattainable in natural social interactions (e.g., Pardo, 2006; Pardo et al., 2010). Yet, it also provides a context a bit closer to natural listening environments than altered vocal feedback (Perkell, 2012; Scheerer & Jones, 2014). The approach does not require explicit perceptual training with feedback to shift category boundaries, as has been used in examining perception-production interactions in sensorimotor adaptation paradigms (Lametti, Krol, Shiller, & Ostry, 2014a). Perhaps most significantly, the Idemaru and Holt paradigm presents the chance to test the open question of whether detailed statistical regularities experienced across dimensions in the acoustic input impact how these dimensions are realized in speech production. To date, most investigations of perception-production interactions examined boundary shifts along single acoustic dimensions (e.g., Babel & Bulatov, 2012; Gentilucci & Bernardis, 2007; Gregory, Dagan, & Webster, 1997; Nielsen, 2011; Shockley et al., 2004; Vallabha & Tuller, 2004) and adjustments to the range of values across these dimensions (e.g., higher/lower F0 frequencies). Using the dynamic adjustments to cue weighting in speech

perception as a tool for investigating perception-production interactions makes it possible to examine more complex acoustic dimension relationships, such as correlations and other distributional regularities among acoustic dimensions.

In the present study, participants heard and categorized one of four words (*beer, pier, deer,* or *tear*) on most trials. Occasionally, participants were prompted visually, without any concurrent acoustic speech information, to say these same words. Unbeknownst to participants, an artificial accent that reversed the F0 x VOT correlation typical of English was introduced in the manner of Idemaru and Holt (2011). If the perceptual effects reported by Idemaru and Holt have an effect on listeners' own speech productions, we expect that speech productions will differ less in F0 across voicing categories in the block with the artificial accent, when F0 is down-weighted perceptually, compared to the blocks that mirror the typical English F0 x VOT regularity. This paradigm allows us to investigate the extent to which perceptual weighting of acoustic dimensions based on the statistical regularities experienced for another voice (here, the correlation of two dimensions) plays a role in calibrating the use of those same acoustic dimensions in one's own speech productions.

Methods

Participants

Twenty-seven monolingual English students from Carnegie Mellon University with normal hearing participated. Two participants were excluded from further analyses for failing to distinguish between the *beer/pier* and *deer/tear* stimuli signaled by unambiguous VOTs as demonstrated by categorization errors greater than 2 standard deviations above the group mean (M =3.76%, SD = 4.16%, cutoff at 12.08% incorrect categorization). Since previous research

with these same stimuli produced near-ceiling categorization performance (Idemaru & Holt, 2011; 2014), these high error rates (16.94% and 13.19% incorrect) suggest noncompliance with the task. Another participant failed to respond during the baseline production condition but completed the other blocks. This participant's data were included in analyses that did not involve baseline production data. This left a total of 25 participants (13 women, 12 men) for the primary analyses; analyses that included the baseline production condition have 24 complete data sets (13 women, 11 men).

Stimuli

The stimuli were from Idemaru and Holt (2011). A female monolingual English talker with Midwest dialect (LLH) recorded multiple citation-form utterances of rhymes *beer, pier, deer,* and *tear* in a sound-isolating booth (22.1 Hz, mono, 16 bit WAV files). Instances of the words were chosen based on recording clarity and roughly equivalent duration. Using these recordings, VOT was cross spliced in seven 10-ms steps to create stimulus series that varied from *beer* to *pier* and from *deer* to *tear* (McMurray & Aslin, 2005). Specifically, the initial consonant burst plus 10-ms increments of the voiceless recordings (*pier, tear*) were spliced onto the voiced recordings (*beer, deer*), replacing an equal duration of the voiced recordings at the beginning of the word. The 0 ms VOT stimuli were created by replacing the voiced burst with the voiceless burst. To create negative VOT durations, 10-ms increments of pre-voicing from voiced recordings was inserted before the burst of the 0 ms VOT stimuli. All splices were made at zero crossings to avoid artifacts. The resulting series ranged from -20-ms to 40-ms VOT for the beer/pier series and -10-ms to 50-ms for the deer/tear series. A stimulus with a perceptually ambiguous VOT value was chosen from each series to serve as the basis for creating Test

stimuli. Consistent with the shift in VOT category boundary across place of articulation (Lisker & Abramson, 1985), this stimulus was 10 ms for *beer/pier* and 20 ms for *deer/tear*. These values were chosen based on pilot testing by Idemaru and Holt (2011).

These two series formed the basis for creating a two-dimensional stimulus grid across which VOT and F0 varied. For each stimulus along each of the VOT series, the original F0 contour of the stimulus was manipulated using Praat 5.0 (Boersma & Weenik, 2013) to adjust the onset F0 of the vowel from 220 Hz to 300 Hz in 10-Hz steps. F0 remained at this frequency for 80 ms, after which it linearly decreased to 180 Hz over 150 ms. This procedure resulted in 126 stimuli (9 steps F0 onset x 7 steps VOT x 2 places of articulation). From this set, 44 stimuli were used in the study (see Fig. 1). These stimuli consisted of Exposure stimuli (open circles, Fig. 1) that clearly indicated one of the four words with a perceptually unambiguous VOT, and Test stimuli (colored circles, Fig. 1) that had an intentionally ambiguous VOT (10 ms for *beer/pier*, 20 ms for *deer/tear*) and either high (290 Hz) or low (230 Hz) F0. These Test stimuli provided a means of measuring the influence of F0 on word recognition when VOT was ambiguous, across blocks.

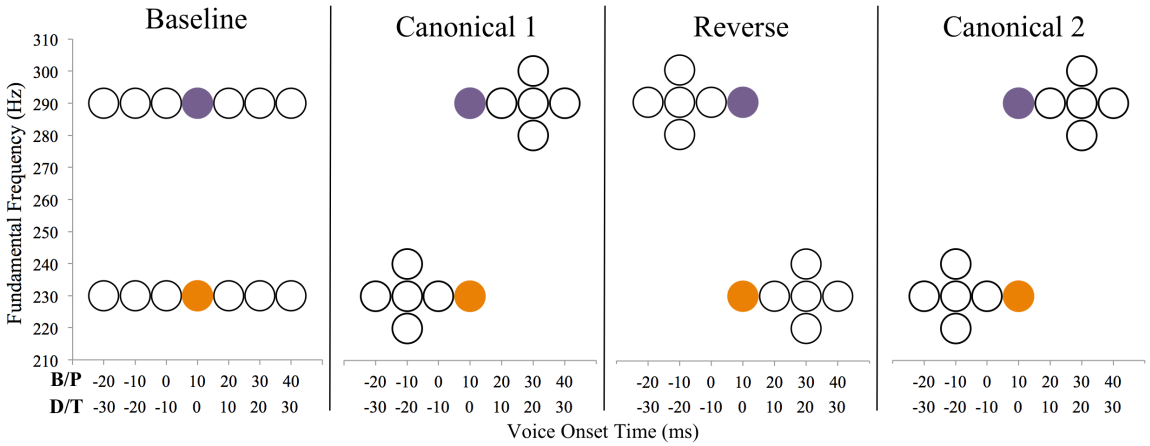

Figure 1. A diagram of stimulus sampling across experiment blocks. Each stimulus is indicated by a symbol in the voice onset time (VOT) by fundamental frequency (F0) acoustic space. Open

symbols show Exposure stimuli whereas colored symbols are Test stimuli. Stimulus sampling varies across blocks to introduce F0 x VOT regularities among Exposure stimuli that are Canonical in relation to long-term experience with English, or that Reverse that relationship.

Procedure

The design was similar to that of Idemaru and Holt (2011). Participants first completed two blocks in which baseline use of F0 and VOT was assessed in perception and production. The order of the perception and production Baseline blocks was counterbalanced across participants. The perception Baseline block included the full range of seven VOT values presented at two different F0 frequencies (230 and 290 Hz) and each place of articulation. This stimulus set included the Test stimuli (colored circles, Fig. 1) that appeared in each of the experiment blocks. Stimuli from the *beer/pier* and *deer/tear* stimulus sets were randomly intermixed. Each stimulus was repeated 5 times, for a total of 140 perceptual trials. On all perceptual trials (both Exposure and Test), participants saw a blank screen for 500 ms, immediately followed by the simultaneous presentation of an acoustic stimulus and visual clip-art images corresponding to the four response choices. The images remained on the screen until the participant responded by pressing a key on the number pad that corresponded to one of the four pictures, indicating the word they heard. The pictures appeared in the same location on each trial and the response keys corresponded to the spatial location of the pictures on the screen.

Participants also completed a Baseline block of picture-naming speech production trials. Production trials were signaled by a visual cue highlighting one of the four icons corresponding to the target words (*beer, pier, deer, tear*) in the absence of any acoustic stimulus. Upon this signal, participants uttered the word once in a 2 second window before the next icon was

highlighted. Spoken words were recorded (22.05 Hz, 16 bit stereo, WAV format) by E-prime 2.0.8.79 (Psychology Software Tools, Pittsburgh, PA) with a Shure SM2 Head-worn microphone connected through a EuroRack UB802 mixing box. The recordings were automatically labeled and digitally stored for later acoustic analysis. Over the course of the production Baseline block each of the words was produced 5 times in randomly ordered sets of four. Average vowel F0 and consonant VOT in voiced versus voiceless productions were measured off-line (see below).

Next, three experimental blocks (Canonical, Reverse, Canonical 2) manipulated the relationship between F0 and VOT in Exposure stimuli (open circles, Fig. 1). All three experimental blocks intermixed voicing and place of articulation, with random presentation order. Throughout these experimental blocks Exposure stimuli possessed perceptually unambiguous VOT values that signaled the identity of the word as *beer* versus *pier* or *deer* versus *tear.* These Exposure stimuli sampled acoustic space such that F0 co-varied with VOT in a manner consistent with long-term regularities of English (short VOT, low F0; Canonical block in Fig. 1) or in a manner that reversed the relationship (short VOT, high F0; Reverse block in Fig. 1) to create an artificial accent. Test trials, for which VOT was perceptually ambiguous (colored circles, Fig. 1), assessed the contribution of F0 as a function of short-term incidental experience with the Exposure trials. Since VOT does not provide information with which to differentiate *beer/pier* and *deer/tear* for Test trials, the extent to which listeners label Test stimuli as voiced versus voiceless consonants is an index of reliance on F0 in word recognition. Test trials were not differentiated from Exposure trials in the experiment. Each of the three experimental blocks consisted of 240 trials (10 Exposure stimuli and 2 Test stimuli for both the *beer/pier* and *deer/tear* stimulus sets and 10 repetitions of each). The stimuli were presented

without breaks or any other overt demarcation; block structure was implicit and unknown to participants.

In contrast to prior studies, trials to elicit participants' speech productions were regularly interspersed throughout the three experimental blocks. After each set of 24 perceptual trials (one of the 10 repetitions), participants were prompted (with a picture, in the same manner as in the Baseline block) to produce a single utterance of each of the four words *(beer/pier/deer/tear*). These speech production trials provided a basis for later acoustic analyses to assess the extent to which exposure to an "artificial accent" affected participants' own use of F0 in distinguishing the target words in their speech productions.

Results

Perception: Word Recognition at Baseline

At baseline, both F0 and VOT influenced voicing categorization (see Fig. 2) as has been reported by many previous studies (Castleman & Diehl, 1996; Chistovich, 1969; Haggard, Summerfield, & Roberts, 1981; Haggard, Ambler, & Callow, 1970; Whalen et al., 1993). Place of articulation errors (e.g., identifying a *beer-pier* stimulus as *deer* or *tear*) were removed from the analysis (1.49% of trials). The proportion of voiceless judgments for the baseline perception block was analyzed for each series (*beer/pier* and *deer/tear*) with a separate 7 (VOT) x 2 (F0) repeated measures ANOVA. For the *beer/pier* series, there was a significant main effect of VOT, $F(6,24) = 528.62$ $p < .001$, $\eta_p^2 = .957$, a main effect of F0, $F(1,24) = 80.83$, $p < .001$, $\eta_p^2 = .771$ and a significant interaction between these factors $F(6,24) = 27.42$ $p < .001$, $\eta_p^2 = .533$.

There was a similar pattern for the *deer/tear* series. Here, we observed a significant main effect of VOT $F(6,24) = 532.69$, $p < .001$, $\eta_p^2 = .957$, a main effect of F0, $F(1,24) = 31.67$, $p <$

.001, $\eta_p^2 = .569$, and a significant interaction between these factors $F(6,24) = 16.58$, $p < .001$, $\eta_p^2 = .409$. These results reflect participants' sensitivity to the long-term regularities of English when tested with balanced exposure to variability across F0 and VOT, and no correlation between the dimensions. The results show that participants used both F0 and VOT in phonetic categorization. Consistent with the relationship between F0 and VOT that characterizes English, the higher-frequency F0 led to more voiceless (*pier, tear*) responses than lower-frequency F0. These results also verify that perceptually ambiguous VOT stimuli chosen as Test stimuli for later segments of the experiment (10 ms for *beer-pier*, 20 ms for *deer-tear*) were in fact ambiguous at baseline, and influenced by F0.

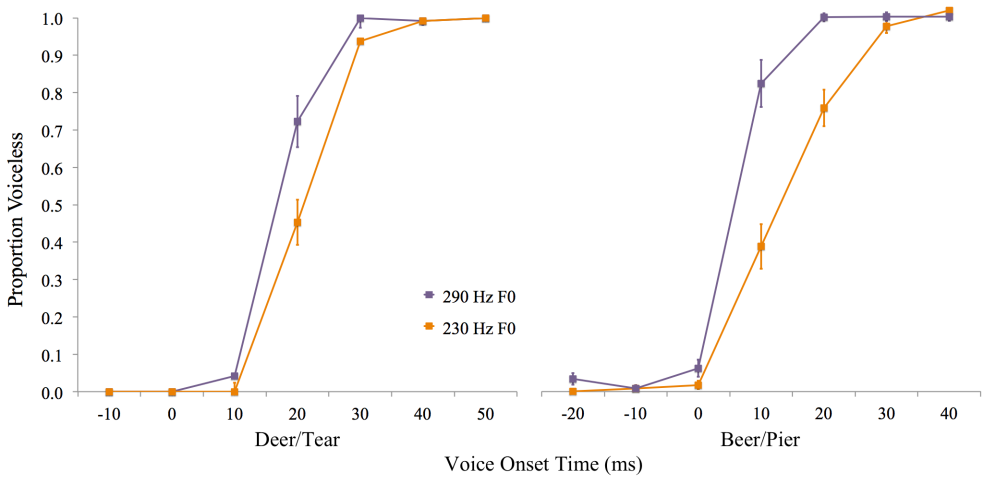

Figure 2. Baseline categorization across the VOT dimension when F0 was high (290 Hz) and low (230 Hz) frequency. The 10 ms VOT stimuli in the *beer-pier* series and the 20 ms VOT stimulus in the *deer-tier* series served as Test stimuli in the later experimental blocks. Error bars show standard error of the mean.

Perception: Word Recognition of Exposure Stimuli

Exposure stimuli were characterized by perceptually unambiguous VOT (see Fig. 1 and

perceptual responses at Baseline in Fig. 2). In Canonical, Reverse and Canonical 2 blocks, listeners reliably used the perceptually unambiguous VOT to accurately recognize the words. After place of articulation errors (based on the series from which each stimulus was constructed) were removed (3.24% of trials), categorization accuracy for Exposure stimuli was near ceiling: the mean proportion of expected (correct, based on VOT values) responses was high for both voiced (*beer* and *deer*), M = .97, *SE* = .01, and voiceless (*pier* and *tear*), M = .94, *SE* = .01, stimuli. This result corroborates the expectation that VOT, as a dimension with strong perceptual weight (Shultz et al., 2012), robustly signals categorization of Exposure trials across the three experimental blocks.

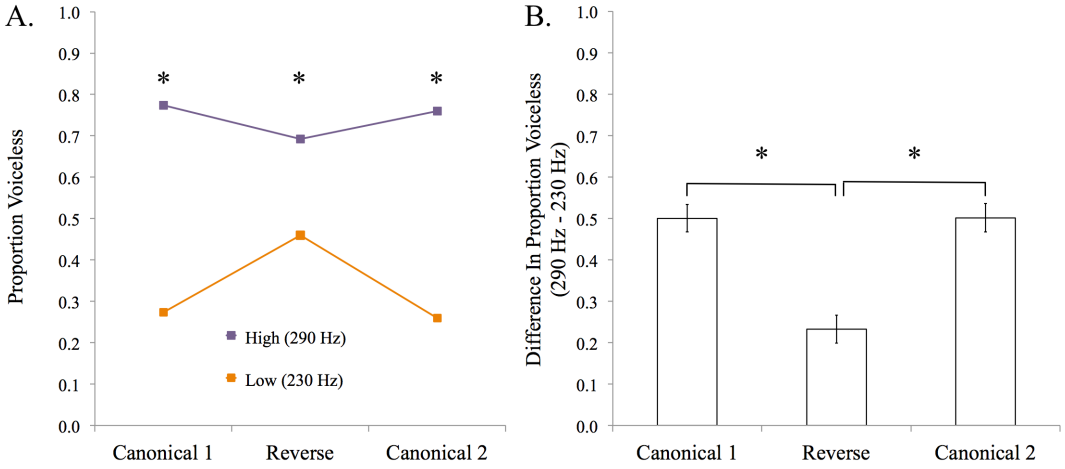

Figure 3. A) The proportion voiceless responses for the high (290 Hz, purple line) and low (230 Hz, orange line) Test stimuli across the three experimental blocks. Asterisks show significant differences in categorization of Test stimuli within each block ($p < .001$). **B)** Difference scores of proportion voiceless responses for high versus low F0 Test stimuli. Error bars represent standard error and asterisks indicate a significant difference ($p < .001$) in the influence of F0 on perceptual categorization across blocks.

Perception: Word Recognition of Test Stimuli

Fig. 3 presents word recognition of the Test stimuli across blocks as proportion of responses categorized as voiceless (*pier, tear*) as a function of high (290 Hz) versus low (230 Hz) frequency F0. Trials for which participants made a place-of-articulation error were excluded from analysis (Canonical: 1.5% of trials; Reverse: 4% of trials; Canonical 2: 1.9% of trials). The remaining data were submitted to a 3 (blocks: Canonical, Reverse, Canonical 2) x 2 (F0: high versus low) repeated measures ANOVA.

A main effect of block, a main effect of F0, and a significant interaction between these factors were observed. The main effect of block, $F(2, 24) = 3.41$, $p = .05$, $\eta_p^2 = .229$, indicates that the overall proportion of voiceless responses varied across the blocks. The main effect of Test stimulus F0 frequency, $F(1, 24) = 218.46$, $p < .001$, $\eta_p^2 = .901$, demonstrates that listeners labeled words with ambiguous VOT more often as voiceless (*pier/tear*) with a high frequency F0 and as voiced (*beer/deer*) with a low frequency F0. In order to examine this main effect of Test stimulus F0, post hoc paired-sample t-tests were run between the high and low F0 Test stimuli within each block. These t-tests showed a consistent effect of F0 on word recognition in each block (see Fig. 3a; Canonical $t(24) = -15.2$, $p < .001$; Reverse $t(24) = -6.87$, $p < .001$; Canonical 2 $t(24) = -14.58$, $p < .001$ (alpha = .017)). Critically, however, the robustness of the influence of F0 on perceptual categorization of the Test stimuli differed across blocks. The significant block x F0 interaction ($F(3, 24) = 46.24$, $p < .001$, $\eta_p^2 = .801$) indicates that the influence of F0 on Test trial categorization was modulated by the short-term deviation in the F0 x VOT correlation experienced across Exposure trials.

The degree of the influence of F0 was examined in more detail by subtracting the proportion of voiceless judgments for the low-F0 Test stimulus from the proportion of voiceless

judgments for the high-F0 Test stimulus within each block. These difference scores were compared using a repeated-measures ANOVA over the three blocks (Canonical, Reverse, Canonical 2). There was a significant main effect of block, $F(2, 24) = 44.15$, $p < .001$, $\eta_p^2 = .648$. The detailed influence of F0 across blocks was compared using post-hoc paired-sample t-tests. Participants relied upon F0 in perceptual categorization more in the Canonical blocks than in the context of exposure to the artificial accent: Canonical-Reverse, $t(24) = 8.94$, $p < .001$; Reverse-Canonical 2, $t(24) = -7.72$, $p < .001$ (alpha =.025). Fig 3b illustrates that the extent to which listeners rely upon F0 in word recognition is diminished in the context of exposure to an artificial accent that reverses the typical F0 x VOT correlation. But, reliance on F0 rebounds when the canonical F0 x VOT relationship is restored in the final canonical block. This replicates the down-weighting in perceptual weight for F0 reported by Idemaru and Holt (2011; 2014; see Liu & Holt, 2015 for a conceptual replication with another speech contrast).

Production: Acoustic Analyses of Speech Production Data

The primary question in the present research is whether this observed pattern of F0 down-weighting in *perception* of another talker's voice likewise affects listeners' reliance on F0 to communicate category distinctions in their own speech productions. To assess this, we analyzed the acoustics of participants' *beer, pier, deer,* and *tear* productions across blocks.

The acoustics of these productions were analyzed from the digital recordings of the productions using Praat version 5.2.26 (Boersma & Weenik, 2013). VOT duration and mean F0 frequency were measured using a combination of Praat scripting and hand labeling. Voiced portions of the production were first identified automatically by converting each sound file to a point process and then into a TextGrid (vuv) with voiced and voiceless portions labeled. The

initial consonant burst was identified by visual inspection, and the VOT was labeled by hand on the TextGrid. At this point, portions of the recordings falsely labeled as voiced by the automated algorithm were corrected. The portion of each vowel labeled as voiced in the TextGrid (vuv) was converted into a Pitch object and mean F0 over this entire voiced portion was measured using the "Get Mean" command in Praat. Trials with negative VOT were set aside and dependent measures were extracted separately, by visual inspection, for these utterances.

Trials for which participants said the wrong word, said more than one word, had a false start, said nothing at all, for which the recording was truncated due to a time-out on the recording period, or for which Praat was unable to calculate F0 were discarded from further analyses (4.6% of all production trials). This left 3338 recordings submitted to analysis (91.6% of Baseline trials, 96.7% of Canonical trials, 97.1% of Reverse trials and 94.2% of trials in the second Canonical block).

Production: Baseline Speech Production Data

We examined participants' baseline reliance on F0 in differentiating voiced from voiceless consonants. Averaged across gender of the participants, the results largely follow the expected pattern of F0 and VOT in English speech (see Tables I and II, for *beer-pier* and *deer-tear*, respectively), with lower F0 frequencies associated with voiced consonants (*beer, deer*) and higher F0 frequencies associated with voiceless consonants (*pier, tear*). It is noted that the relationship was smaller and less reliable for *beer-pier* than *deer-tear*. This is quite interesting because the influence of F0 on voicing categorization, its perceptual weight, is consistently stronger for the /d/-/t/ contrast than for the /b/-/p/ contrast in baseline perceptual assessments

(Idemaru & Holt, 2011; 2014). If this sample is representative, it suggests an alignment of perceptual weight with subtle differences in English speech productions.

F0 varies substantially across gender (e.g., Peterson & Barney, 1952), so we also analyzed F0 as a function of voiced and voiceless consonants separately for male and female participants. At baseline, female participants used F0 to signal voicing for *beer-pier* (see Table III) and for *deer-tear* (see Table IV). Interestingly, the male participants did not use F0 to contrast *beer* from *pier* (see Table III) and used it only marginally to contrast *deer-tear* (Table IV). Overall, males used F0 less to differentiate between voiced and voiceless productions at baseline. This may relate to prior studies reporting greater F0 dynamics among females (Babel, 2012; Babel & Bulatov, 2012; Chang, 2012). We discuss this unexpected result in the General Discussion.

Table I. Descriptive statistics and t-test results for VOT and mean F0 in baseline *beer* and *pier* productions.

Acoustic Dimension	Beer		Pier		Mean Difference		95% CI of the Mean Difference	t(23)	*p*
	M	*SE*	*M*	*SE*	*M*	*SE*			
VOT	15.37	1.97	78.03	4.17	-62.66	3.74	[-70.4, -54.91]	-16.73	<.001
Mean F0	172.03	10.89	177.02	12.09	-4.98	2.44	[-10.03, .08]	-2.04	.053

Table II. Descriptive statistics and t-test results for VOT and mean F0 in baseline *deer* and *tear* productions.

Acoustic Dimension	Deer		Tear		Mean Difference		95% CI of the Mean Difference	t(23)	*p*
	M	*SE*	*M*	*SE*	*M*	*SE*			
VOT	22.20	1.74	87.60	4.05	-65.40	3.87	[-73.41, -57.39]	-16.89	<.001
Mean F0	168.85	11.30	175.48	11.67	-6.62	2.06	[-10.88, -2.37]	-3.23	.004

Table III. Descriptive statistics and t-test results for mean F0 in baseline *beer* and *pier* productions, split by gender.

Acoustic Dimension	Beer		Pier		Mean Difference		95% CI of the Mean Difference	t	df	*p*
	M	*SE*	*M*	*SE*	*M*	*SE*				
Male F0	121.18	4.37	121.81	4.84	-.64	2.98	[-7.27, 6.00]	-0.21	10	.835
Female F0	215.15	8.36	223.62	10.13	-8.46	3.51	[-1.27, -2.57]	-2.41	12	.033

Table IV. Descriptive statistics and t-test results for mean F0 in baseline *deer* and *tear* productions, split by gender.

Acoustic Dimension	Deer		Tear		Mean Difference		95% CI of the Mean Difference	t	df	*p*
	M	*SE*	*M*	*SE*	*M*	*SE*				
Male F0	117.18	4.34	121.82	4.81	-4.64	2.17	[-9.46, .19]	-2.14	10	.058
Female F0	212.69	9.51	221.15	9.51	-8.46	3.30	[-15.65, -1.27]	-2.57	12	.025

Production: Analysis of VOT Across Experimental Blocks

In the present study, the manipulation of dimensions is such that VOT persists as a consistent signal to category membership, even as the relationship to F0 varies with the introduction of the artificial accent in the Reverse block. Prior research demonstrates that VOT perceptual weight is not influenced by introduction of the accent (Idemaru & Holt, 2011; 2014). As a result, VOT serves as an excellent control condition for speech production analyses. We expect no significant differences in speakers' use of VOT across blocks.

To test this prediction, VOT durations extracted from *beer-pier* and *deer-tear* speech productions were independently submitted to 3 (block; Canonical, Reverse, Canonical 2) x 2

(voicing category: voiced, voiceless) repeated-measures ANOVAs. Significant main effects of voicing for both *beer-pier*, $F(1,24) = 262.14$, $p < .001$, $\eta_p^2 = .916$, and *deer-tear*, $F(1,24) = 285.83$, $p < .001$, $\eta_p^2 = .923$, indicate that participants used VOT to contrast the voicing categories, as expected. There was no main effect of block for *beer-pier* $F(2,23)=1.42$, $p = .262$, $\eta_p^2 = .11$ or *deer-tear* $F(2,23)=1.96$, $p = .163$, $\eta_p^2 = .146$ and no interaction of block and voicing category for *beer-pier* $F(2,23)=.10$, $p = .899$, $\eta_p^2 = .009$, or *deer-tear* $F(2,23)=2.15$, $p = .139$, $\eta_p^2 = .158$. This demonstrates that participants used VOT consistently in speech production, even as the introduction of the artificial accent in the Reverse block changed the typical relationship of F0 and VOT and led to perceptual down-weighting of F0. This consistency in participants' VOT across blocks also assures that any changes observed in reliance on F0 in speech productions across blocks is unlikely to arise from participant fatigue or measurement error.

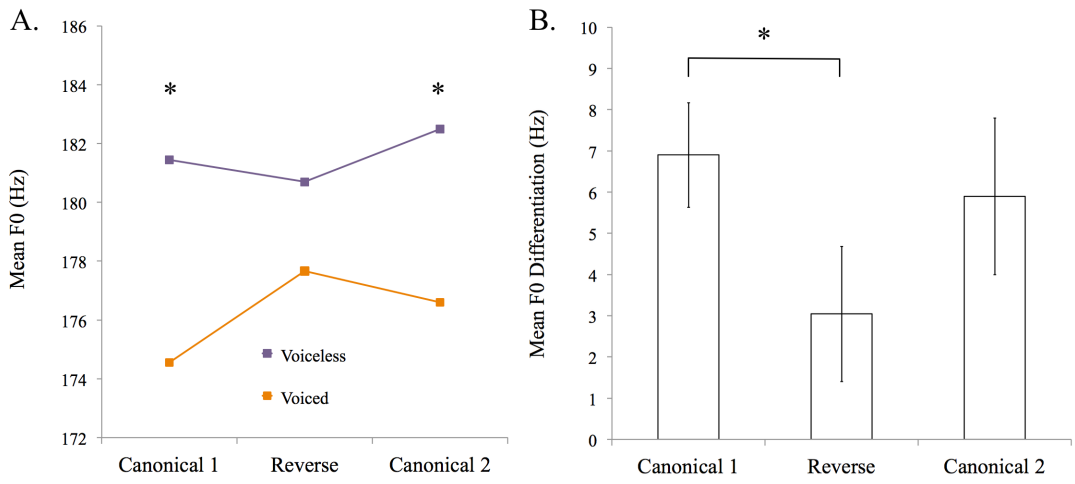

Figure 4. A) Mean F0 measured from utterances of voiced and voiceless words across Canonical, Reverse, and Canonical 2 blocks. Asterisks indicate significant differences in participants' use of F0 in distinguishing voiced versus voiceless in their own speech productions ($p<=.005$). **B)** Mean F0 differences between voiced and voiceless productions across the experimental blocks. Error bars are standard error of the mean and the asterisk indicates a

significant difference in reliance on F0 to differentiate voiced vs. voiceless productions across blocks (p=.015).

Production: Analysis of F0 Across Experimental Blocks

The goal of the current research was to examine whether the perceptual reweighting of F0 in response to the artificial accent in the Reverse block affects how participants use F0 in producing the same words. Decreased reliance on F0 to signal voiced versus voiceless productions in the Reverse block relative to the Canonical blocks would present strong evidence that acoustic dimension re-weighting in perception exerts an effect on the use of those same dimensions in production. To test this prediction, we compared the difference in mean F0[1] for voiced and voiceless productions across experimental blocks.

Fig. 4 plots the mean F0 of all participants' voiced (*beer* and *deer*) and voiceless (*pier* and *tear*) speech productions as a function of the experimental blocks. It is important to note that despite the visual similarity in patterning of the perception and production data (see Fig. 3 versus Fig. 4), the production and perception analyses rely upon different dependent measures. Perceptual down-weighting of F0 in response to exposure to the artificial accent is reflected as a reduced differentiation of Test stimuli with perceptually ambiguous VOT, but distinct F0s, in perceptual categorization. In contrast, the coincident down-weighting of F0 in listeners' own speech productions would be reflected as the difference in *mean F0* for voiced (*beer/deer*) versus voiceless (*pier/tear*) speech productions.

The mean F0s from participants' voiced and voiceless speech productions were entered into a 3 (block: Canonical, Reverse, Canonical 2) x 2 (voicing: voiced, voiceless) repeated

[1] Mean F0 was chosen based on a lower variability in this measure relative to other measures of F0 such as initial F0 or F0 over the first 50 ms.

measures ANOVA. Most important to the predictions of the present study, there was a significant interaction effect between block and voicing with multivariate tests[2] $F(2,23) = 3.52$, $p = .046$, $\eta_p^2 = .234$. The interaction indicates that participants' reliance on F0 to signal voiced and voiceless categories in their own productions differed across blocks. We explored the nature of this interaction with post hoc paired sample t-tests comparing F0 frequencies for voiced and voiceless productions within each block. There were significant differences in F0 for each Canonical block (Canonical: $t(24) = -5.44$, $p < .001$; Canonical 2: $t(24) = -3.1$, $p = .005$; alpha =.017). However, F0 did not differentiate voiced and voiceless productions in the Reverse block, $t(24) = -1.86$, $p = .08$. This pattern of results indicates that exposure to an artificial accent in another talker's voice had an effect on listeners' expression of one of the acoustic dimensions in their own speech productions (see Fig. 4a). When the experienced relationship between F0 and VOT was consistent with long-term regularities of English in the Canonical blocks, participants used both VOT (see above) and F0 to differentiate voiced and voiceless productions. Exposure to the reversed correlation of F0 and VOT in the Reverse block reduced participants' reliance on the F0 dimension in differentiating voiced and voiceless speech productions. In addition to the interaction, there was a main effect of voicing $F(1,24) = 17.48$, $p < .001$, $\eta_p^2 = .421$ with participants producing lower frequency F0s for vowels following voiced consonants (M = 176.27 Hz, SE = 10.63 Hz) than voiceless consonants (M = 181.55 Hz, SE = 11.26 Hz) across blocks. There was no main effect of block, $F(2,23)=.47$, $p = .629$, $\eta_p^2 = .04$ indicating that the global F0 across all (voiced and voiceless) productions did not change systematically across experimental blocks.

[2] A multivariate test was used in this mixed model ANOVA because the univariate approach presumes sphericity of the error variance-covariance matrix. The sphericity assumption was not met in the present data, $W(2,23)=.74$, $p=.03$ (Mauchy's test of sphericity) suggesting the appropriateness of a multivariate approach (Maxwell & Delaney, 2004).

Difference scores between the mean vowel F0 of voiced and voiceless productions were analyzed using a repeated measures ANOVA over the three blocks (Canonical, Reverse, Canonical 2). There was a main effect of block $F(2,23) = 3.52$, $p = .046$, $\eta_p^2 = .234$ indicating that incidental experience with shifting F0 x VOT correlations across the perceptual Exposure stimuli affected participants' use of F0 to differentiate voiced and voiceless productions (see Fig. 4b). Post hoc *t*-tests showed that this was driven by the reduced reliance on F0 to differentiate voiced and voiceless productions in the Reverse block (M = 3.04 Hz, SE = 1.64 Hz) relative to the first Canonical block (M = 6.90 Hz, SE = 1.27 Hz), $t(24) = 2.63$, $p = .015$. There was no significant difference in reliance on F0 in the final Canonical block (M = 5.90 Hz, SE = 1.90) compared to the Reverse block, $t(24) = -1.32$, $p = .2$ (alpha = .025). This significant change in the use of F0 to differentiate voiced from voiceless productions indicates that the short term perceptual exposure that influences the acoustic dimension weighting of F0 in perception also influences the use of F0 in the interleaved productions of the same words.

Whereas some studies have identified significant correlations between the observed changes in speech motor learning and perceptual adaptation (Lametti, Krol, Shiller, & Ostry, 2014a; Nasir & Ostry, 2009) other studies have not observed correlations (Lametti, Rochet-Capellan, Neufeld, Shiller, & Ostry, 2014b; Shiller et al., 2009; see Cressman & Henriques, 2009 in in visuomotor adaptation). In the present study, there was no correlation between individual participant's degree of perceptual down-weighting and the change in their reliance on F0 to distinguish voicing categories in their own speech across the Canonical and Reverse blocks ($r = .14$, $p = .493$). The lack of correlation in perception and production cue weights may not be surprising in light of several studies demonstrating that individuals do not necessarily rely upon the same cues in perception that they do in production (Idemaru & Holt, 2013; Schertz, Cho,

Lotto, & Warner, 2015; Shultz et al., 2012). Consistent with other studies, we observe adaptation in speech production as a result of exposure to perceptual regularities. At a group level, these perceptual weights are mirrored in speech production. However, there was no correlation in individual's *degree* of down-weighting in perception and production.

General Discussion

The basis of the relationship between speech perception and production remains unresolved despite a long history of investigation (e.g., Fowler, 1986; Guenther & Vladusich, 2012; Gupta & MacWhinney, 1997; Hickok, Houde, & Rong, 2011; Hume & Johnson, 2001; Liberman & Mattingly, 1985; MacDonald, 2013; Perkell, 2012; Pickering & Garrod, 2013). Yet, the effects of heard speech on one's own speech productions are observed both in fairly natural listening environments (Chang, 2012; Pardo et al., 2012) and in tightly controlled experimental conditions with artificial acoustic manipulations (Fowler et al., 2003; Lametti, Krol, Shiller, & Ostry, 2014a). The current research demonstrates for the first time that the distributional regularities between acoustic dimensions experienced incidentally in the perception of another talker's voice exert an influence on speech production in a rapid and quickly-reversible manner that affects both perception and production of the detailed acoustic dimensions that signal speech categories.

In the present study, listeners simply recognized words from a small set (*beer, pier, deer, tear*) varying in onset consonant, with occasional visual prompts to produce these same words. Unbeknownst to participants, the subtle relationship between two acoustic dimensions (F0 and VOT) conveying the onset consonants varied across blocks, thereby creating an "artificial accent." In accord with prior research examining dimension-based statistical learning (Idemaru

& Holt, 2011; 2014; Liu & Holt, 2015), perceptual reliance on F0 rapidly adapted in response to exposure to the artificial accent. As participants categorized speech with acoustic dimensional regularities that matched native English experience (Canonical blocks), they reliably differentiated between Test stimuli with perceptually-ambiguous VOT, using only F0. When perceptual exposure shifted so that the correlation between F0 and VOT was reversed relative to canonical English regularities (Reverse block), F0 was no longer as effective in signaling category membership; it was perceptually down-weighted. It is important to note that these perceptual shifts in the cue weighting for F0 occurred incidentally as participants simply recognized the words. There was no overt training as in prior studies investigating perception - production interactions (Lametti, Krol, Shiller, & Ostry, 2014a; Shiller et al., 2013). However, as in prior perceptual research (Idemaru & Holt, 2011; 2014; Liu & Holt, 2015), the dynamic changes in cue weighting in response to the artificial accent cannot be understood as a broad direction of attention away from the down-weighted F0 dimension because listeners quickly returned to rely on F0 as a signal for consonant identity when the canonical F0 x VOT correlation was reinstated in the final Canonical block. This indicates that listeners continue to track F0, even as its influence on consonant categorization is down-weighted in the Reverse block.

In the present study, we discovered for the first time that this dimension-based statistical learning evokes a corresponding influence on listeners' own speech productions. As the relationship between F0 and VOT in another talker's voice changed over the course of the experiment, listeners' reliance on these dimensions in their own speech productions was affected. In the Reverse block in which listeners down-weighted perceptual reliance upon F0 in consonant categorization, listeners' own productions of these consonants were less differentiated by F0.

Importantly, this change in speech production was specific to F0, the dimension down-weighted in perception in the Reverse block. In contrast, VOT use remained consistent across blocks. Further, the variability of F0 (as shown by standard error of the difference scores in the analysis of F0 productions across experimental blocks) is quite stable, suggesting that the results cannot be attributed simply to more variable use of F0 across the experiment. The adjustment to the detailed acoustic dimensions in listeners' own speech occurred without any explicit training, overt speech shadowing (there was no acoustic model on production trials), or modified auditory vocal feedback. This demonstrates that short-term deviations in the perceptual relationships among acoustic dimensions, such as those that might occur in natural speaking and listening environments, can subtly shift the use of those acoustic dimensions in both perception *and* production.

At a computational level of analysis directed at *what* the system does (Marr & Poggio, 1976), one can understand these effects through Bayesian models in which the likelihood of perceiving a phonetic category depends upon distributions of cues associated with the category in prior experience (Kleinschmidt & Jaeger, 2015). Bayesian accounts model adaptive plasticity in speech perception through incremental adjustments to these cue distributions based on recent experience. These adjustments shift the cue distributions, resulting in phonetic category boundary shifts of the sort observed for lexically- and visually-mediated perceptual speech adaptation (Kleinschmidt & Jaeger, 2015). Although these models have not been applied directly to the down-weighting perceptual cue weights examined in the present study, the Bayesian principles are compatible with adjustments of perception based on long-term regularities by short-term deviations in the input. Bayesian models provide an elegant means of understanding the computational demands of adaptation in speech perception. However, they do not speak to

how the system implements these computations, the representations upon which they act, or the nature of the processing involved.

We contend that understanding 'the *how*' at an algorithmic level of explanation (Marr & Poggio, 1976) will be significant in developing a complete model of speech adaptation and its interaction with speech production. Bayesian models may be compatible with many algorithmic implementations since they depict the nature of computations involved in a phenomenon like adaptive plasticity in speech perception. Algorithmic models attempt to specify the implementation of these computations. As such, the two can be wholly compatible. As advocated by Marr, each will be important in understanding a phenomenon. Nonetheless, from our perspective, a central challenge lies in understanding the detailed nature of the cognitive processes and representations involved. Specifying how adaptive plasticity is implemented in online speech perception will be essential to understanding the fingerprints it leaves on speech production.

One proposal of 'the *how'* is that a general, supervised error-driven learning mechanism that adjusts the mapping from acoustic input to pre-lexical representations can account for the shifts in perceptual cue weighting observed in dimension-based statistical learning (Guediche, Blumstein, Fiez, & Holt, 2014a; Idemaru & Holt, 2011; Liu & Holt, 2015). The proposal is that perceptually unambiguous information from the input, such as the unambiguous VOT in the present Exposure stimuli, will be sufficient to activate phonetic categories based on strong acoustic-to-category mappings established by long-term experience. As a consequence of category activation, the system may generate a prediction of the expected information along other acoustic dimensions associated with long-term experience. When predictions differ from the actual input (as in the case of F0 in the Reverse block Exposure stimuli) the discrepancy may

result in an internally-generated error signal that can drive adaptive adjustments of the internal prediction to improve alignment of future predictions with the incoming input. In the present study, these adaptive adjustments lead to decreased perceptual weighting of F0.

Such supervised error-driven learning can be implemented through a connectionist model, as recently described (Liu and Holt, 2015). Beginning from interactive activation models like TRACE (McClelland & Elman, 1986; Mirman, McClelland, & Holt, 2006) whereby long-term mappings between input and pre-lexical representations are realized as connection weights among network representations, learning may be implemented to adjust the efficiency of connection weights. Indeed, this approach accounts for other forms of adaptation in speech perception when the connection weights are adjusted via Hebbian learning (Mirman et al., 2006). However, Hebbian learning may be too sluggish to accommodate the rapid learning that emerges with limited exposure to the artificial accent (Guediche et al., 2014a; Idemaru & Holt, 2014). In this regard, supervised learning mechanisms may be better aligned with the rapid time course of learning observed in the present results (see Bertelson, Vroomen, & de Gelder, 2003; Guediche et al., 2014a; Guediche, Holt, Laurent, Lim, & Fiez, 2014b; Idemaru & Holt, 2011; Liu & Holt, 2015; Norris, McQueen, & Cutler, 2003; Vroomen, van Linden, de Gelder, & Bertelson, 2007).

In order for this mechanism to explain the present results, error-driven learning in perception must influence production, a proposal that is parsimonious with contemporary models of perception-production interactions in speech (e.g., Guenther & Vladusich, 2012; Gupta & MacWhinney, 1997; Hickok, 2012). For example, the "auditory error map" proposed by the Directions into Velocities of Articulators (DIVA) model can communicate perceptual changes (potentially those induced by dimension-based statistical learning) to the articulator position and velocity maps that are proposed to drive speech output (Guenther & Vladusich, 2012). Similarly,

adjustments in the auditory syllable targets induced through dimension-based statistical learning ultimately might affect the motor syllable programs proposed by the hierarchical state feedback control model (Hickok, 2012). The neural locus for where these error signals in perception exert their influence on production differs between models, however the underlying mechanism (error-driven learning) is common across models. Therefore, error signals generated by a discrepancy between expected and experienced acoustic input affecting the articulatory system seems a plausible explanation for the current results – a conclusion also reached by recent research into perceptual influences on speech production (Bourguignon et al., 2016). Error-driven supervised learning provides a potential mechanism by which dimension-based statistical learning may occur in the empirical perceptual findings thus far, and the current results suggest that error signals generated in the perceptual system influence productions as well – a finding compatible with models of speech production.

In the sense that unambiguous information can drive activation of established representations (such as phonetic categories) and thereby generate predictions that may be compared against the incoming input, Guediche et al., (2014) have suggested that adaptive plasticity of speech perception evident in dimension based statistical learning may share commonalities with adaptive plasticity observed in other short-term adjustments in speech perception, such as lexically-mediated perceptual learning (Kraljic & Samuel, 2005; 2006; Norris et al., 2003; Samuel & Kraljic, 2009). In lexically-mediated perceptual learning paradigms, participants are exposed to words with an acoustically-ambiguous speech sound embedded in a lexical context that disambiguates its identity. For example, participants might hear an acoustically-ambiguous fricative that falls between /s/ and /ʃ/ (hence, /~sʃ/) embedded in words consistent with /s/ (e.g., *ambiguous*) whereas other participants hear the same /~sʃ/ sound

in /ʃ/-consistent words (e.g., *abolish*). Following exposure to lexically-disambiguating contexts like this, participants who experience /s/-consistent lexical contexts categorize non-word syllables that include the ambiguous fricative more often as /s/ than participants who experience /~sʃ/ in the context of /ʃ/-consistent words (Kraljic & Samuel, 2005).

For both dimension-based statistical learning and lexically-mediated perceptual learning, acoustic information that deviates from long-term expectations about native speech is disambiguated by additional information (lexical context in the case of lexically-mediated perceptual learning, the unambiguous VOT dimension in the present study). Over the course of experience with the disambiguating information, there is rapid learning that calibrates subsequent speech perception. Even when the disambiguating context is no longer available to support perception, the calibration persists for the shifted speech category. In the case of Kraljic and Samuel (2005), this learning is evidenced by a category boundary shift across groups of participants who experience /s/-consistent versus /ʃ/-consistent experience with /~sʃ/. In the present study, it was observed as perceptual down-weighting of an acoustic dimension, F0. As noted above, Guediche et al. (2014) make a case that there may be common supervised error-driven learning mechanisms supporting learning in each paradigm. As yet, this remains unresolved by the empirical literature. However, it does suggest the utility of comparing outcomes across these somewhat different experimental paradigms.

This is especially relevant because the impact of lexically-mediated perceptual learning on listeners' own speech productions has been investigated. Kraljic, Brennan, and Samuels (2008) attempted to understand how lexically-mediated perceptual learning across /s/-/ʃ/ might influence listeners' own speech productions. They examined two lexical conditions that disambiguated /~sʃ/. In one condition, the unusual acoustics of the perceptually-ambiguous /~sʃ/

could be attributed to a regional dialect (it could be resolved to be /s/, but only when it occurred before /tr/, in keeping with speech typical of Long Island) and another condition for which the ambiguous token occurred in place of each /s/ token, without regard to context. Each condition was experienced by a different group of participants and each group was familiar with the Long Island dialect. Speech productions of words containing /s/, /ʃ/, and /str/ were elicited before and after the lexically-mediated perceptual learning. The group that could attribute the unusual acoustics of /~sʃ/ to the Long Island dialect due to the context-selectivity of the experience with /~sʃ/ with /tr/ context showed no perceptual learning. However, the group that experienced ambiguous /~sʃ/ across all /s/-consistent words did exhibit perceptual learning (Kraljic, Brennan, & Samuel, 2008). This finding suggests that lexically-mediated perceptual learning is sensitive to cognitive attributions of the distorted acoustic signal. However, most germane to the present study, it is notable that there were no effects on speech production for either group.

There are a variety of factors that may contribute to the different pattern of results observed by Kraljic et al. (2008), as compared to the present study. Kraljic et al. (2008) examined the hypothesis that the spectral mean of fricative productions would shift as a function of perceptual learning by measuring acoustics across 24 words (8 with /s/, 8 with /ʃ/, and 8 with /str/). In contrast, we assessed the impact of learning on F0 across just 4 words (2 voiced, 2 voiceless). It is possible that repetition of a smaller word corpus reduced acoustic variability, increasing our sensitivity to detect an effect on speech production. Perhaps relatedly, the Kraljic et al. study had fewer repetitions of each word (2 repetitions in the pre-test and 2 in the post-test) than the present study (10 repetitions of each word interspersed throughout each block). Another possibility is that our focus on F0 as an acoustic dimension across which to measure the impact on speech production was propitious. F0 has been shown to be especially prominent in effects of

adaptation on speech production (Babel & Bulatov, 2012; Gregory et al., 1997). It is possible that it is easier to detect a shift in F0 than a shift in the spectral mean of fricative productions. Of course, it is also possible that the learning involved across these different paradigms involves different modes of interaction with speech production. Future research that systematically examines the impact of various sources of adaptive plasticity in speech perception on speech production will be essential in understanding the detailed nature of perception-production interactions.

An additional question for future investigation is whether talker identity plays a role in the perception-production interactions we observe here. In the present study, the effects were driven by exposure to a female talker in a mixed-gender sample of listeners. It is compelling that we observe an effect of the female talker on listeners' own speech productions since the F0 range of male and female listeners differs quite dramatically (see Tables I and II) and our male participants failed to use F0 reliably to differentiate their own voicing categories in the baseline block. Previous results have observed gender differences in shadowing (Babel, 2012; Babel & Bulatov, 2012; Namy, Nygaard, & Sauerteig, 2002), accommodation (Lelong & Bailly, 2011), and responses to foreign language immersion (Chang, 2012) whereby female listeners and speakers generally exhibit stronger adaptation to input than their male counterparts. Indeed, the social environment, including factors such as gender, may have broad influence. Interlocutors often show convergence in their speech over the course of experimentally controlled conversations (Garrod & Anderson, 1987; Levitan & Hirschberg, 2011; Pardo, 2006), and social factors such as attractiveness (Babel, 2012), race (Babel, 2012), gender (Lelong & Bailly, 2011; Pardo et al., 2010; Van Bezooijen, 1995), and social role (Pardo et al., 2010; 2013) of interlocutors have short-term influences on speech productions. Stereotypes about gender also

influence speech perception (Strand, 1999; Strand & Johnson, 1996). Collectively, this suggests the possibility that social expectations and experimentally manipulated social factors may influence short term, dynamic changes in acoustic dimension use across production and perception. The role of such social factors (or attributions, as in Kraljic et al., 2008) in guiding adaptive changes makes sense given that expectations during speech perception are often shaped by these factors. If adaptation is driven by deviations from expectations (as predicted by error-driven supervised learning, and compatible with Bayesian accounts) then social factors may provide critical contextual information that establishes these expectations. Although the current experiment was not designed to address these issues, the present paradigm sets the stage to investigate the details of how the model talker, social factors, participant gender, and baseline acoustic dimension weighting interact to produce interactions between speech perception and production. An advantage of the present approach is that it allows for controlled manipulation of acoustic regularities in the perceptual input and measurement of the detailed acoustic dimensions upon which those regularities depend in speech production while maintaining fairly natural, and incidental, conditions for learning.

The present results establish that perception-production interactions exist in dimension-based statistical learning and demonstrate that statistical regularities between dimensions in the input affect speech production. Although further research will be required to fully understand the nature of this perception-production interaction, past studies of the perceptual bases of dimension-based statistical learning provide some insights to guide the work. The stimuli in the present experiment were all lexical items (*beer, pier, deer, tear*). Since lexical information provided no information with which to disambiguate perception (and thus no lexical "teaching signal," in contrast to lexically-mediated perceptual learning paradigms, Norris et al., 2003;

Kraljic & Samuels, 2005), it seems likely that the site of learning is pre-lexical. This is supported by recent findings from Liu and Holt (2015) whereby dimension-based statistical learning is observed in perception of non-words and, further, learning across non-words generalizes to words not heard in the artificial accent. Perhaps even more compelling, Lehet and Holt (2015) find that dimension-based statistical learning does not impact early encoding at the acoustic dimension level. Even when a dimension is perceptually down-weighted in response to an artificial accent as in the present study, the dimension nonetheless maintains its ability to influence perception of subsequent speech through perceptual contrast (Lehet & Holt, 2015). Jointly, these studies present both bottom-up and top-down constraints. Given these constraints, it appears likely that exposure to the artificial accent influences the weighted mappings from acoustic dimensions to phonetic categories (see Liu & Holt, 2015; Idemaru & Holt, 2014 for discussion). The same constraints on the perceptual data also constrain the nature of perception-production interactions that give rise to the present results.

The present paradigm provides an excellent test-bed for advancing this understanding. For example, Idemaru and Holt (2014) find that when the F0 x VOT correlations for /b/-/p/ (Canonical, Reverse, Canonical) and /d/-/t/ (Reverse, Canonical, Reverse) are opposing within experiment blocks, listeners do not sum the statistics across "voicing" (which would result in a null F0 x VOT relationship). Instead, listeners independently track F0 x VOT correlations for /b/-/p/ and /d/-/t/ and dimension-based statistical learning follows accordingly. Listeners down-weight F0 for /d/-/t/ when the artificial accent reverses the F0 x VOT relationship even as they maintain reliance on F0 for /b/-/p/ that maintains the canonical English F0 x VOT relationship in the same block. Investigations of the information that the system uses to determine the "bins" into which regularities are calculated promise to inform the nature of representations involved in

early speech processing, and interactions with speech production. Such studies can help to refine the representations somewhat loosely described in current neurobiological models of speech perception-production interactions.

A rich and growing literature of studies on the adaptive plasticity in speech perception indicates that rapid learning can adjust perception of speech that deviates from the norms established by long-term regularities of the language community, thereby aiding speech perception in adverse listening conditions (for reviews see Guediche et al., 2014a; Mattys, Davis, Bradlow, & Scott, 2012). Various teaching signals, including visual (Bertelson et al., 2003; Vroomen et al., 2007), phonotactic (Cutler, McQueen, & Butterfield, 2008), lexical (Kraljic et al., 2008; Kraljic & Samuel, 2005; Norris et al., 2003) and statistical (Clayards, Tanenhaus, Aslin, & Jacobs, 2008; Idemaru & Holt, 2011; 2014) information can guide these rapid adjustments. The present results demonstrate that this learning has a concomitant influence on speech production.

References

Babel, M. (2012). Evidence for phonetic and social selectivity in spontaneous phonetic imitation. *Journal of Phonetics*, *40*(1), 177–189. http://doi.org/10.1016/j.wocn.2011.09.001

Babel, M., & Bulatov, D. (2012). The Role of Fundamental Frequency in Phonetic Accommodation. *Language and Speech*, *55*(2), 231–248. http://doi.org/10.1177/0023830911417695

Bertelson, P., Vroomen, J., & de Gelder, B. (2003). Visual recalibration of auditory speech identification: a McGurk after effect. *Psychological Science*, *14*(6), 592–597.

http://doi.org/10.1046/j.0956-7976.2003.psci_1470.x

Boersma, P., & Weenik, D. (2013). Praat: Doing phonetics by computer (Version 5.3. 39) Computer program]. Retrieved January 29, 2013.

Bourguignon, N. J., Baum, S. R., & Shiller, D. M. (2016). Please Say What This Word Is—Vowel-Extrinsic Normalization in the Sensorimotor Control of Speech. *Journal of Experimental Psychology: Human Perception and Performance*, 1–10. http://doi.org/10.1037/xhp0000209

Castleman, W. A., & Diehl, R. L. (1996). Effects of Fundamental Frequency on Medial and Final [Voice] Judgments. *Journal of Phonetics*, *24*(4), 383–398.

Chang, C. B. (2012). Rapid and multifaceted effects of second-language learning on first-language speech production. *Journal of Phonetics*, *40*(2), 249–268. http://doi.org/10.1016/j.wocn.2011.10.007

Chang, C. B. (2013). A novelty effect in phonetic drift of the native language. *Journal of Phonetics*, *41*(6), 520–533. http://doi.org/10.1016/j.wocn.2013.09.006

Chistovich, L. A. (1969). Variation of the fundamental voice pitch as a discriminatory cue for consonants. *Soviet Physics and Acoustics*, *14*, 372–378.

Clayards, M., Tanenhaus, M. K., Aslin, R. N., & Jacobs, R. A. (2008). Perception of speech reflects optimal use of probabilistic speech cues. *Cognition*, *108*(3), 804–809. http://doi.org/10.1016/j.cognition.2008.04.004

Coupland, N., & Giles, H. (1988). Introduction the communicative contexts of accommodation. *Language & Communication*, *8*(3-4), 175–182. http://doi.org/10.1016/0271-5309(88)90015-8

Cressman, E. K., & Henriques, D. Y. P. (2009). Sensory Recalibration of Hand Position

Following Visuomotor Adaptation. *Journal of Neurophysiology*, *102*(6), 3505–3518. http://doi.org/10.1152/jn.00514.2009

Cutler, A., McQueen, J. M., & Butterfield, S. (2008). Prelexically-driven perceptual retuning of phoneme boundaries (p. 2056). Presented at the Proceedings of Interspeech.

Delvaux, V., & Soquet, A. (2007). The influence of ambient speech on adult speech productions through unintentional imitation. *Phonetica-Basel*.

Diehl, R. L., & Kingston, J. (1991). Phonetic covariation as auditory enhancement: The case of the [+ voice]/[-voice] distinction (14 ed., pp. 139–143). Current phonetic research paradigms: Implications for speech motor control, PERILUS.

Donath, T. M., Natke, U., & Kalveram, K. T. (2002). Effects of frequency-shifted auditory feedback on voice F0 contours in syllables. *The Journal of the Acoustical Society of America*, *111*(1), 357–366.

Fowler, C. A. (1986). An event approach to the study of speech perception from a direct-realist perspective. *Journal of Phonetics*, *14*(1), 3–28.

Fowler, C., Brown, J. M., Sabadini, L., & Welhing, J. (2003). Rapid access to speech gestures in perception: Evidence from choice and simple response time tasks. *Journal of Memory and Language*, *49*(3), 396–413. http://doi.org/10.1016/S0749-596X(03)00072-X

Francis, A. L., Baldwin, K., & Nusbaum, H. C. (2000). Effects of training on attention to acoustic cues. *Perception & Psychophysics*, *62*(8), 1668–1680. http://doi.org/10.3758/BF03212164

Francis, A. L., Kaganovich, N., & Driscoll-Huber, C. (2008). Cue-specific effects of categorization training on the relative weighting of acoustic cues to consonant voicing in English. *The Journal of the Acoustical Society of America*, *124*(2), 1234.

http://doi.org/10.1121/1.2945161

Garrod, S., & Anderson, A. (1987). Saying what you mean in dialogue: A study in conceptual and semantic co-ordination. *Cognition*, *27*(2), 181–218. http://doi.org/10.1016/0010-0277(87)90018-7

Gentilucci, M., & Bernardis, P. (2007). Imitation during phoneme production. *Neuropsychologia*, *45*(3), 608–615. http://doi.org/10.1016/j.neuropsychologia.2006.04.004

Giles, H. (1973). Accent mobility: A model and some data. *Anthropological Linguistics*, 87–105. http://doi.org/10.2307/30029508

Giles, H. (1977). Social psychology and applied linguistics: Towards an integrative approach (Vol. 33, pp. 27–42). ITL: Review of applied linguistics.

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review*, *105*(2), 251–279. http://doi.org/10.1037/0033-295X.105.2.251

Gregory, S. W., Jr, Dagan, K., & Webster, S. (1997). Evaluating the relation of vocal accommodation in conversation partners' fundamental frequencies to perceptions of communication quality. *Journal of Nonverbal Behavior*, *21*(1), 23–43. http://doi.org/10.1023/A:1024995717773

Guediche, S., Blumstein, S. E., Fiez, J., & Holt, L. L. (2014a). Speech perception under adverse conditions: insights from behavioral, computational, and neuroscience research. *Frontiers in Systems Neuroscience*, *7*, 1–16. http://doi.org/10.3389/fnsys.2013.00126/abstract

Guediche, S., Holt, L. L., Laurent, P., Lim, S.-J., & Fiez, J. (2014b). Evidence for cerebellar contributions to adaptive plasticity in speech perception. *Cerebral Cortex*, (bht428).

Guenther, F. H., & Vladusich, T. (2012). A neural theory of speech acquisition and production. *Journal of Neurolinguistics*, *25*(5), 408–422. http://doi.org/10.1016/j.jneuroling.2009.08.006

Guion, S. (2002). The vowel systems of Quichua-Spanish bilinguals. Age of acquisition effects on the mutual influence of the first and second languages. *Phonetica*, *60*(2), 98–128. http://doi.org/10.1159/000071449

Gupta, P., & MacWhinney, B. (1997). Vocabulary Acquisition and Verbal Short-Term Memory: Computational and Neural Bases. *Brain and Language*, *59*(2), 267–333. http://doi.org/10.1006/brln.1997.1819

Haggard, M. P., Summerfield, Q., & Roberts, M. (1981). Psychoacoustical and cultural determinants of phoneme boundaries: Evidence from trading F_0 cues in the voiced–voiceless distinction. *Journal of Phonetics*, *9*(1), 49–62.

Haggard, M., Ambler, S., & Callow, M. (1970). Pitch as a Voicing Cue. *Journal of the Acoustical Society of America*, *47*(2), 613–617.

Hazan, V., & Barrett, S. (2000). The development of phonemic categorization in children aged 6–12. *Journal of Phonetics*, *28*(4), 377–396. http://doi.org/10.1006/jpho.2000.0121

Hickok, G. (2012). Computational neuroanatomy of speech production. *Nature Reviews Neuroscience*, *13*(2), 135–145. http://doi.org/10.1038/nrn3158

Hickok, G., Houde, J., & Rong, F. (2011). Sensorimotor Integration in Speech Processing: Computational Basis and Neural Organization. *Neuron*, *69*(3), 407–422. http://doi.org/10.1016/j.neuron.2011.01.019

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*(5), 3059–3071. http://doi.org/10.1121/1.2188377

Honorof, D. N., Weihing, J., & Fowler, C. A. (2011). Articulatory events are imitated under rapid shadowing. *Journal of Phonetics*, *39*(1), 18–38.

http://doi.org/10.1016/j.wocn.2010.10.007

Houde, J. F., & Jordan, M. I. (1998). Sensorimotor adaptation in speech production. *Science*, *279*(5354), 1213–1216. http://doi.org/10.1126/science.279.5354.1213

Hume, E., & Johnson, K. (2001). A model of the interplay of speech perception and phonology. *Working Papers in Linguistics-Ohio State University Department of Linguistics*.

Idemaru, K., & Holt, L. L. (2011). Word recognition reflects dimension-based statistical learning. *Journal of Experimental Psychology: Human Perception and Performance*, *37*(6), 1939–1956. http://doi.org/10.1037/a0025641

Idemaru, K., & Holt, L. L. (2013). The developmental trajectory of children's perception and production of English /r/-/l/. *The Journal of the Acoustical Society of America*, *133*(6), 4232. http://doi.org/10.1121/1.4802905

Idemaru, K., & Holt, L. L. (2014). Specificity of dimension-based statistical learning in word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, *40*(3), 1009–1021. http://doi.org/10.1037/a0035269

Idemaru, K., Holt, L. L., & Seltman, H. (2012). Individual differences in cue weights are stable across time: The case of Japanese stop lengths. *The Journal of the Acoustical Society of America*, *132*(6), 3950. http://doi.org/10.1121/1.4765076

Ingvalson, E. M., McClelland, J. L., & Holt, L. L. (2011). Predicting native English-like performance by native Japanese speakers. *Journal of Phonetics*, *39*(4), 571–584. http://doi.org/10.1016/j.wocn.2011.03.003

Iverson, P., & Kuhl, P. K. (1995). Mapping the perceptual magnet effect for speech using signal detection theory and multidimensional scaling. *Journal of the Acoustical Society of America*, *97*(1), 553–562.

Iverson, P., Kuhl, P. K., Akahane-Yamada, R., Diesch, E., Tohkura, Y., Kettermann, A., & Siebert, C. (2003). A perceptual interference account of acquisition difficulties for non-native phonemes. *Cognition*, *87*(1), B47–B57. http://doi.org/10.1016/S0010-0277(02)00198-1

Kingston, J., & Diehl, R. L. (1994). Phonetic Knowledge. *Language*, *70*(3), 419. http://doi.org/10.2307/416481

Kleinschmidt, D. F., & Jaeger, T. F. (2015). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review*, *122*(2), 148–203. http://doi.org/10.1037/a0038695

Kohler, K. J. (1982). F0 in the Production of Lenis and Fortis Plosives. *Phonetica*, *39*(4-5), 199–218. http://doi.org/10.1159/000261663

Kohler, K. J. (1984). Phonetic Explanation in Phonology: The Feature Fortis/Lenis. *Phonetica*, *41*(3), 150–174. http://doi.org/10.1159/000261721

Kraljic, T., & Samuel, A. G. (2005). Perceptual learning for speech: Is there a return to normal? *Cognitive Psychology*, *51*(2), 141–178. http://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*(2), 262–268. http://doi.org/10.3758/BF03193841

Kraljic, T., Brennan, S. E., & Samuel, A. G. (2008). Accommodating variation: Dialects, idiolects, and speech processing. *Cognition*, *107*(1), 54–81. http://doi.org/10.1016/j.cognition.2007.07.013

Lametti, D. R., Krol, S. A., Shiller, D. M., & Ostry, D. J. (2014a). Brief Periods of Auditory Perceptual Training Can Determine the Sensory Targets of Speech Motor Learning. *Psychological Science*, *25*(7), 0956797614529978–1336.

http://doi.org/10.1177/0956797614529978

Lametti, D. R., Rochet-Capellan, A., Neufeld, E., Shiller, D. M., & Ostry, D. J. (2014b). Plasticity in the Human Speech Motor System Drives Changes in Speech Perception. *Journal of Neuroscience*, *34*(31), 10339–10346. http://doi.org/10.1523/JNEUROSCI.0108-14.2014

Lehet, M., & Holt, L. L. (2015). Adaptation to accent affects categorization, but not basic perceptual representation. *The Journal of the Acoustical Society of America*, *137*(4), 2384–2384. http://doi.org/10.1121/1.4920674

Lelong, A., & Bailly, G. (2011). Study of the Phenomenon of Phonetic Convergence Thanks to Speech Dominoes. In *Analysis of Verbal and Nonverbal Communication and Enactment. The Processing Issues* (Vol. 6800, pp. 273–286). Berlin, Heidelberg: Springer Berlin Heidelberg. http://doi.org/10.1007/978-3-642-25775-9_26

Levitan, R., & Hirschberg, J. B. (2011). Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions (pp. 3081–3084). Presented at the INTERSPEECH, Florence, Italy.

Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition*, *21*(1), 1–36.

Lisker, L. (1986). "Voicing" in English: A Catalogue of Acoustic Features Signaling /b/ Versus /p/ in Trochees. *Language and Speech*, *29*(1), 3–11.

Lisker, L., & Abramson, A. S. (1985). Relative Power of Cues: F0 Shift Versus Voice Timing. In V. Fromkin (Ed.), *Phonetic Linguistics* (pp. 25–33). Academic Press Inc.

Liu, R., & Holt, L. L. (2015). Dimension-based statistical learning of vowels. *Journal of Experimental Psychology: Human Perception and Performance*, *41*(6), 1783–1798.

http://doi.org/10.1037/xhp0000092

Lord, G. (2008). Second language acquisition and first language phonological modification. In J. Bruhn do Garavito & E. Valenzuela (Eds.), (pp. 184–193). Presented at the Selected proceedings of the 10th hispanic linguistics symposium, Somerville, MA.

Lotto, A. J., Sato, M., & Diehl, R. L. (2004). Mapping the task for the second language learner: the case of Japanese acquisition of/r/and/l. In J. Slifka, S. Manuel, & M. Matthies (Eds.), (pp. C181–C186). Presented at the From sound to sense: 50+ Years of Discoveries in Speech Communication.

Lowenstein, J. H., & Nittrouer, S. (2008). Patterns of acquisition of native voice onset time in English-learning children. *The Journal of the Acoustical Society of America*, *124*(2), 1180. http://doi.org/10.1121/1.2945118

MacDonald, M. C. (2013). How language production shapes language form and comprehension. *Frontiers in Psychology*, *4*, 226. http://doi.org/10.3389/fpsyg.2013.00226

Marr, D., & Poggio, T. (1976). From Understanding Computation to Understanding Neural Circuitry.

Marslen-Wilson, W. (1973). Linguistic structure and speech shadowing at very short latencies. *Nature*, *244*(5417), 522–523. http://doi.org/10.1038/244522a0

Mattys, S. L., Davis, M. H., Bradlow, A. R., & Scott, S. K. (2012). Speech recognition in adverse conditions: A review. *Language and Cognitive Processes*, *27*(7-8), 953–978. http://doi.org/10.1080/01690965.2012.705006

Maxwell, S. E., & Delaney, H. D. (2004). Designing Experiments and Analyzing Data (Vol. 1). Psychology Press.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive*

Psychology, *18*(1), 1–86. http://doi.org/10.1016/0010-0285(86)90015-0

McMurray, B., & Aslin, R. N. (2005). Infants are sensitive to within-category variation in speech perception. *Cognition*, *95*(2), B15–B26. http://doi.org/10.1016/j.cognition.2004.07.005

Miller, R. M., Sanchez, K., & Rosenblum, L. D. (2013). Is speech alignment to talkers or tasks? *Attention, Perception & Psychophysics*, *75*(8), 1817–1826. http://doi.org/10.3758/s13414-013-0517-y

Mirman, D., McClelland, J. L., & Holt, L. L. (2006). An interactive Hebbian account of lexically guided tuning of speech perception. *Psychonomic Bulletin & Review*, *13*(6), 958–965.

Namy, L. L., Nygaard, L. C., & Sauerteig, D. (2002). Gender Differences in Vocal Accommodation: The Role of Perception. *Journal of Language and Social Psychology*, *21*(4), 422–432. http://doi.org/10.1177/026192702237958

Nasir, S. M., & Ostry, D. J. (2009). Auditory plasticity and speech motor learning. *Proceedings of the National Academy of Sciences*, *106*(48), 20470–20475.

Nielsen, K. (2011). Specificity and abstractness of VOT imitation. *Journal of Phonetics*, *39*(2), 132–142. http://doi.org/10.1016/j.wocn.2010.12.007

Nittrouer, S. (2004). The role of temporal and dynamic signal components in the perception of syllable-final stop voicing by children and adults. *The Journal of the Acoustical Society of America*, *115*(4), 1777–1790. http://doi.org/10.1121/1.1651192

Nittrouer, S., Lowenstein, J. H., & Packer, R. R. (2009). Children discover the spectral skeletons in their native language before the amplitude envelopes. *Journal of Experimental Psychology: Human Perception and Performance*, *35*(4), 1245–1253. http://doi.org/10.1037/a0015020

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive*

Psychology, *47*(2), 204–238. http://doi.org/10.1016/S0010-0285(03)00006-9

Pardo, J. S. (2006). On phonetic convergence during conversational interaction. *The Journal of the Acoustical Society of America*, *119*(4), 2382–2393. http://doi.org/10.1121/1.2178720

Pardo, J. S. (2013). Measuring phonetic convergence in speech production. *Frontiers in Psychology*, *4*, 559. http://doi.org/10.3389/fpsyg.2013.00559

Pardo, J. S., Gibbons, R., Suppes, A., & Krauss, R. M. (2012). Phonetic convergence in college roommates. *Journal of Phonetics*, *40*(1), 190–197. http://doi.org/10.1016/j.wocn.2011.10.001

Pardo, J. S., Jay, I. C., & Krauss, R. M. (2010). Conversational role influences speech imitation. *Attention, Perception & Psychophysics*, *72*(8), 2254–2264. http://doi.org/10.3758/BF03196699

Pardo, J. S., Jay, I. C., Hoshino, R., Hasbun, S. M., Sowemimo-Coker, C., & Krauss, R. M. (2013). Influence of Role-Switching on Phonetic Convergence in Conversation. *Dx.Doi.org*, *50*(4), 276–300. http://doi.org/10.1080/0163853X.2013.778168

Perkell, J. S. (2012). Movement goals and feedback and feedforward control mechanisms in speech production. *Journal of Neurolinguistics*, *25*(5), 382–407. http://doi.org/10.1016/j.jneuroling.2010.02.011

Peterson, G. E., & Barney, H. L. (1952). Control Methods Used in a Study of the Vowels. *The Journal of the Acoustical Society of America*, *24*(2), 175–184.

Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioral and Brain Sciences*, *36*(04), 329–347. http://doi.org/10.1017/S0140525X12001495

Purcell, D. W., & Munhall, K. G. (2006). Compensation following real-time manipulation of

formants in isolated vowels. *The Journal of the Acoustical Society of America*, *119*(4), 2288. http://doi.org/10.1121/1.2173514

Roon, K. D., & Gafos, A. I. (2014). Perceptuo-motor effects of response-distractor compatibility in speech: beyond phonemic identity. *Psychonomic Bulletin & Review*, *22*(1), 242–250. http://doi.org/10.3758/s13423-014-0666-6

Samuel, A. G., & Kraljic, T. (2009). Perceptual learning for speech. *Attention, Perception & Psychophysics*, *71*(6), 1207–1218. http://doi.org/10.3758/APP.71.6.1207

Sancier, M. L., & Fowler, C. A. (1997). Gestural drift in a bilingual speaker of Brazilian Portuguese and English. *Journal of Phonetics*, *25*(4), 421–436. http://doi.org/10.1006/jpho.1997.0051

Scheerer, N. E., & Jones, J. A. (2014). The predictability of frequency-altered auditory feedback changes the weighting of feedback and feedforward input for speech motor control. *European Journal of Neuroscience*, *40*(12), 3793–3806. http://doi.org/10.1111/ejn.12734

Schertz, J., Cho, T., Lotto, A., & Warner, N. (2015). Individual differences in phonetic cue use in production and perception of a non-native sound contrast. *Journal of Phonetics*, *52*, 183–204. http://doi.org/10.1016/j.wocn.2015.07.003

Shiller, D. M., Sato, M., Gracco, V. L., & Baum, S. R. (2009). Perceptual recalibration of speech sounds following speech motor learning. *The Journal of the Acoustical Society of America*, *125*(2), 1103. http://doi.org/10.1121/1.3058638

Shiller, D., Lametti, D., & Ostry, D. J. (2013). Auditory plasticity and sensorimotor learning in speech production (pp. 060150–060150). Presented at the ICA 2013 Montreal, ASA. http://doi.org/10.1121/1.4799848

Shockley, K., Sabadini, L., & Fowler, C. A. (2004). Imitation in shadowing words. *Perception &*

Psychophysics, *66*(3), 422–429. http://doi.org/10.3758/BF03194890

Shultz, A. A., Francis, A. L., & Llanos, F. (2012). Differential cue weighting in perception and production of consonant voicing. *The Journal of the Acoustical Society of America*, *132*(2), EL95. http://doi.org/10.1121/1.4736711

Strand, E. A. (1999). Uncovering the Role of Gender Stereotypes in Speech Perception. *Journal of Language and Social Psychology*, *18*(1), 86–100. http://doi.org/10.1177/0261927X99018001006

Strand, E. A., & Johnson, K. (1996). Gradient and visual speaker normalization in the perception of fricatives. In D. Gibbon (Ed.), *Natural Language Processing and Speech Technology, Results of the 3rd KONVENS Conference* (pp. 14–26).

Vallabha, G. K., & Tuller, B. (2004). Perceptuomotor bias in the imitation of steady-state vowels. *The Journal of the Acoustical Society of America*, *116*(2), 1184–1197. http://doi.org/10.1121/1.1764832

Van Bezooijen, R. (1995). Sociocultural Aspects of Pitch Differences between Japanese and Dutch Women. *Language and Speech*, *38*(3), 253–265. http://doi.org/10.1177/002383099503800303

Villacorta, V. M., Perkell, J. S., & Guenther, F. H. (2007). Sensorimotor adaptation to feedback perturbations of vowel acoustics and its relation to perception. *The Journal of the Acoustical Society of America*, *122*(4), 2306. http://doi.org/10.1121/1.2773966

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*(3), 572–577. http://doi.org/10.1016/j.neuropsychologia.2006.01.031

Whalen, D. H., Abramson, A. S., Lisker, L., & Mody, M. (1993). F0 gives voicing information

even with unambiguous voice onset times. *The Journal of the Acoustical Society of America*, *93*(4), 2152–2159. http://doi.org/10.1121/1.406678

Yamada, R. A., & Tohkura, Y. (1992). The effects of experimental variables on the perception of American English /r/ and /l/ by Japanese listeners. *Perception & Psychophysics*, *52*(4), 376–392. http://doi.org/10.3758/BF03206698