COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 45 (2021) e12947

ISSN: 1551-6709 online
DOI: 10.1111/cogs.12947

The Learning Signal in Perceptual Tuning of Speech: Bottom Up Versus Top-Down Information

Xujin Zhang, Yunan Charles Wu, Lori L. Holt

Department of Psychology, Carnegie Mellon University

Received 14 April 2017; received in revised form 4 January 2021; accepted 5 January 2021

Abstract

Cognitive systems face a tension between stability and plasticity. The maintenance of long-term representations that reflect the global regularities of the environment is often at odds with pressure to flexibly adjust to short-term input regularities that may deviate from the norm. This tension is abundantly clear in speech communication when talkers with accents or dialects produce input that deviates from a listener's language community norms. Prior research demonstrates that when bottom-up acoustic information or top-down word knowledge is available to disambiguate speech input, there is short-term adaptive plasticity such that subsequent speech perception is shifted even in the absence of the disambiguating information. Although such effects are well-documented, it is not yet known whether bottom-up and top-down resolution of ambiguity may operate through common processes, or how these information sources may interact in guiding the adaptive plasticity of speech perception. The present study investigates the joint contributions of bottom-up information from the acoustic signal and top-down information from lexical knowledge in the adaptive plasticity of speech categorization according to short-term input regularities. The results implicate speech category activation, whether from top-down or bottom-up sources, in driving rapid adjustment of listeners' reliance on acoustic dimensions in speech categorization. Broadly, this pattern of perception is consistent with dynamic mapping of input to category representations that is flexibly tuned according to interactive processing accommodating both lexical knowledge and idiosyncrasies of the acoustic input.

Keywords: Speech perception; Adaptive plasticity; Lexically guided phonetic tuning; Dimension-based statistical learning

1. Introduction

Cognitive systems, whether biological or artificial, confront a dilemma in the balance between stability and plasticity (McCloskey & Cohen, 1989; Ratcliff, 1990). Systems

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

must remain plastic enough to accommodate new short-term information, but not be so flexible as to overwrite accumulated long-term knowledge. Speech communication presents an ecologically significant example of the tension between stability and plasticity in cognitive systems, more generally.

On the side of stability, adult listeners have established speech representations that reflect the long-term distributional regularities present among the acoustic dimensions that signal speech categories in a particular language community (Francis, Kaganovich, & Driscoll-Huber, 2008; Holt & Lotto, 2006; Idemaru, Holt, & Seltman, 2012; Iverson et al., 2003; Kondaurova & Francis, 2008; Toscano & McMurray, 2010). The speech categories /b/ and /p/ provide an example. Although both voice onset time (VOT) and fundamental frequency (F0) contribute to signaling /b/ versus /p/, VOT is a more reliable predictor of the categories in English speech productions and, therefore, it more strongly predicts listeners' categorization responses than F0. Correspondingly, VOT carries greater perceptual weight in categorization. Yet listeners do rely upon the secondary, F0, dimension in a manner that respects the fact that English speakers tend to produce /p/ with a somewhat higher F0 than /b/. Accordingly, speech tokens with a perceptually ambiguous VOT tend to be categorized as /p/ when F0 is higher, but as /b/ when F0 is lower (Kingston & Diehl, 1994; Kohler, 1982, 1984). Listeners also can utilize top-down information such as lexical knowledge to aid speech categorization. A speech token is more likely to be categorized as an alternative that completes a real English word, especially when acoustic speech input is acoustically ambiguous. For example, an utterance with an ambiguous VOT may be categorized as /b/ in __*eef* context, but as /p/ in __*eace* context (Ganong, 1980).

Nonetheless, the system remains plastic and can accommodate the fact that the speech we encounter often does not necessarily match exactly the long-term distributional regularities that adults acquired through language development. Talker differences, speech impairments, dialects, accents, and other factors systematically influence the acoustic regularities present in speech input, and can alter the relationship of acoustic speech input to linguistically relevant speech representations in the short term. Thus, speech communication involves more than just learning long-term regularities across speech input as they relate to linguistically relevant representations. It also involves the flexibility to adjust when short-term speech regularities depart from patterns typical of the long-term experiences that established the mappings, using information from both bottom-up and top-down sources.

Indeed, speech perception exhibits adaptive plasticity and rapidly adjusts when top-down knowledge is available to resolve acoustic ambiguities. A rich literature demonstrates the adaptive manner by which speech categorization is "tuned" by short-term experience with lexical knowledge that departs from the norm (Guediche, Blumstein, Fiez, & Holt, 2014; Idemaru & Holt, 2011; Mattys, Davis, Bradlow, & Scott, 2012; Norris, McQueen, & Cutler, 2003; Samuel & Kraljic, 2009; Schwab, Nusbaum, & Pisoni, 1985; Vroomen, van Linden, de Gelder, & Bertelson, 2007). For example, when ambiguous speech is repeatedly resolved by lexical knowledge (e.g., /b/ in __*eef* context), there is rapid *lexically driven perceptual learning* that shifts speech categorization such that the

ambiguous speech is more likely to be categorized as the word-consistent alternative. This rapid tuning is thought to originate from effects of knowledge on pre-lexical processing, although the exact mechanism is debated (Guediche et al., 2014; Kleinschmidt & Jaeger, 2015; McClelland, Mirman, & Holt, 2006; Mirman, McClelland, & Holt, 2006; Norris et al., 2003).

Likewise, low-level information such as acoustic dimensions with strong perceptual weight in signaling speech categories also can drive rapid adaptive plasticity in speech perception. When short-term regularities between dimensions (e.g., like the typical correlation between VOT and F0 in English) deviate from long-term norms, there is rapid re-weighting of the effectiveness of acoustic dimensions in signaling speech categories (Idemaru & Holt, 2011, 2014; Lehet & Holt, 2017, 2020; Liu & Holt, 2015; Schertz, Cho, Lotto, & Warner, 2016; Zhang & Holt, 2018). For example, when listeners encounter an "artificial accent" that reverses the F0 × VOT correlation typical of English, the diagnosticity of F0 in /b/-/p/ categorization is rapidly down-weighted—F0 is much less effective in signaling speech category membership as /b/ versus /p/. This *acoustically driven perceptual learning* has been argued to arise when unambiguous bottom-up acoustic information (e.g., VOT) is available to resolve phonetic category membership and drive adjustment of the effectiveness of secondary acoustic dimensions to speech representations *without* employing lexical knowledge (Idemaru & Holt, 2011; Liu & Holt, 2015).

Acoustically and lexically driven adaptive plasticity have been investigated independently using distinct behavioral paradigms (Eisner & McQueen, 2005; Idemaru & Holt, 2011; Liu & Holt, 2015; Norris et al., 2003; Samuel & Kraljic, 2009). Of course, outside the laboratory, speech input tends to provide both acoustic and lexical information, each of which could support adaptive plasticity in speech perception. Beyond moving toward conditions that capture the information available in natural speech input, merging investigation of how bottom-up and top-down information sources drive adaptive plasticity in speech perception can advance understanding of the means by which speech processing manages the tension between stability and plasticity. If, for example, top-down lexical and bottom-up acoustic information influence different levels of speech processing, they may fail to produce adaptive plasticity effects that align in a common paradigm. Alternatively, these distinct information sources may exert their influence at the same level and produce qualitatively similar adaptive plasticity effects.

The current study examines contributions of both lexical and acoustic information to adaptive plasticity in speech perception in a common paradigm in order to better understand the mechanisms involved. We take interactive activation of levels of representation as a starting point (McClelland & Elman, 1986), positing that bottom-up acoustic information through the primary acoustic dimension and top-down lexical knowledge achieve the same effect of activating phonetic category representation(s) consistent with the information they convey. Therefore, we hypothesize that selective activation of phonetic categories, whether by bottom-up acoustic input or top-down lexical knowledge, will be sufficient to drive adaptive plasticity of the effectiveness of acoustic dimensions in speech categorization. Said another way, we posit that these distinct information sources will be able to exert a common influence, through activation of the phonetic category

representation, on adaptive adjustments in speech perception. We thus expect the pattern of adaptive plasticity to be similar across top-down and bottom-up information. If, instead, they elicit distinct patterns of adaptive plasticity, it would call into question our assumption that the lexical and acoustic information activate the same category representation, or our hypothesis that category activation drives adaptive plasticity.

In the present study, we test this hypothesis by manipulating lexical context such that speech categorization is biased toward /b/ (e.g., __*eef* context, for which *beef* is a word, but *peef* is not) or /p/ (__*eace* context for which *peace* is a word, but *beace* is not) and the presence or absence of perceptually unambiguous bottom-up acoustic information available for speech categorization (i.e., VOT). This approach provides a direct test of whether top-down resolution of phonetic categories through lexical knowledge is sufficient to drive tuning of the influence of acoustic dimensions in speech categorization. Further, it moves investigations forward in examining the joint influence of acoustic and lexical information sources investigated independently in prior research. Just as important, the study has the potential to inform the hotly debated issue of whether lexical activation impacts phonetic processing directly through interactive processing, or at a post-perceptual decision stage through feedforward processing (e.g., McClelland et al., 2006; Norris et al., 2000).

2. Methods

2.1. Participants

Twenty-six native English monolinguals with self-reported normal hearing participated. Volunteers were recruited from the Carnegie Mellon University and randomly assigned to one of two conditions that differed only in the order of tasks.

2.2. Stimuli

A monolingual native-English female adult speaker (L.L.H.) digitally recorded multiple repetitions of the words and nonwords shown in Table 1 in a sound-attenuated booth (44.1 kHz sampling frequency). The tokens were spoken in isolation in citation form. From these recordings, we chose a single token of *beash* and a single token of *peash* based on the clarity of recording, and the tokens' approximately equivalent duration. These exemplars served as endpoints from which to create a stimulus series that varied from /biʃ/ and /piʃ/ (*beash-peash*). These nonwords were chosen for their lexically neutral context and also for the ease in extracting the final fricative from the initial consonant-vowel.

From these natural speech tokens, we first extracted the /bi/-/pi/ consonant-vowel segment from the /biʃ/ and /piʃ/ tokens by removing the portion of the waveform from onset to offset of the consonant /ʃ/ at zero-crossings. We then created a common /bi/-/pi/ series varying in VOT that would serve as the initial consonant-vowel of each of the stimulus classes shown in Table 1. Following the approach of Francis et al. (2008), we edited the stimulus waveforms of the natural speech tokens to create a nine-step series varying in VOT from −20 to 40 ms. The steps were sampled in 10 ms from −20 to 0 ms, then in

Table 1
Stimulus types. There were four stimulus spaces varying in fundamental frequency (F0) and voice onset time (VOT) to create stimuli that varied perceptually from /b/ to /p/. All stimuli began with identical initial consonant-vowel syllables heard as /bi/ or /pi/. The final consonant varied (/ʃ/, /k/, /f/, /s/) to create word and nonword contexts, as shown

	/b/	/p/
Nonword-Nonword (NW-NW)	*beash*, /biʃ/	*peash*, /piʃ/
Word-Word (W-W)	*beak*, /bik/	*peak*, /pik/
Word-Nonword (W-NW)	*beef*, /bif/	*peef*, /pik/
Nonword-Word (NW-W)	*beace*, /bis/	*peace*, /pis/

5 ms from 0 to 20 ms, and again in 10 ms from 20 to 40 ms. This approach provided a fine-grained sampling of perceptually ambiguous VOT tokens (5–15 ms), with less sampling resolution for tokens expected to be perceptually unambiguous. The first 10 ms of the original voiceless (*peash*) production was left intact to preserve the consonant burst. From this starting point in the waveform, 10-ms (or 5-ms) segments (with minor variability so that edits were made at zero-crossings) were excised from the waveform using Praat 5.0 (Boersma & Weenink, 2017), thereby creating stimuli with incrementally shorter VOTs. For the negative VOT values, prevoicing was taken from voiced productions of the same speaker and inserted before the burst in durations varying from −20 to 0 ms, in 10-ms steps.

Returning to the original set of natural utterances recorded by the native-English talker, we extracted the final /k/ from an instance of *beak* (/bik/), a final /f/ from *beef* (/bif/), and a final /s/ from *peace* (/pis/). Each of these final consonants was appended to the waveforms of each stimulus comprising the nine-step /bi/-/pi/ series. As shown in Table 1, this resulted in a word-word (W-W) *beak-peak* series (630 ms), a word-nonword (W-NW) *beef-peef* series (650 ms), and a nonword-word (NW-W) *beace-peace* series (630 ms).

We then manipulated the fundamental frequency (F0) of each series so that the F0 onset frequency of the vowel, /i/, following the word-initial stop consonant was adjusted from 220 to 300 Hz in 10-Hz steps. For each stimulus, the F0 contour of the original production was measured and manually manipulated using Praat 5.0 (Boersma & Weenink, 2017) to adjust the target onset F0. The F0 remained at the target frequency for the first 80 ms of the vowel; from there, it linearly decreased over 150 ms to 180 Hz. This resulted in three 2-dimensional F0 × VOT acoustic spaces across *beace-peace* (NW-W), *beef-peef* (W-NW), and *beak-peak* (W-W), whereby stimuli varied across nine steps along the acoustic VOT dimension and nine steps along the acoustic F0 dimension.

2.3. Procedure

2.3.1. Overview

Participants were seated in front of a computer monitor in a sound-attenuated booth. Each trial involved presentation of a single spoken utterance presented diotically over headphones (Beyer DT-150) and response options presented on the monitor. The position

of response choices was counterbalanced across participants but was consistent across trials for an individual participant. On each trial participants responded to indicate the word or nonword they had heard by pressing a keyboard key corresponding to the orthographic (or picture) label's screen position. The experiment was completed in a single 1-h session across which E-prime (Psychology Software Tools, Inc.) controlled sound presentation, timing, and response collection.

All participants completed each block of each experimental condition after completing an acoustic pretest to establish baseline interactions of F0 and VOT in a lexically neutral context and then a lexical pretest to assess the influence of lexical knowledge and F0 in lexically biased contexts. These pretests served to demonstrate that the acoustic and lexical information manipulated across the experimental conditions do indeed resolve perceptual ambiguity in speech input.

Next, participants completed three experimental conditions (acoustic, lexical, and acoustic + lexical), each with two blocks of trials. For each condition, one block possessed short-term input regularities aligned with English (canonical), whereas the other (reverse) reversed these regularities to create an "artificial accent." This was accomplished across exposure trials that comprised 90% of trials in a block. Across experimental conditions, exposure trials were indicated by bottom-up acoustic information (an unambiguous VOT), top-down lexical information (word knowledge), or a combination of acoustic + lexical information (unambiguous VOT and word knowledge). The remaining 10% of trials were test trials that provided a measure of the extent to which F0 contributed to speech categorization within the block. These trials were identical across blocks and experimental conditions. The test trial stimuli possessed a perceptually ambiguous VOT and neutral lexical information (*beak-peak*, both words) and varied only in F0. In this way, differences in /b/-/p/ categorization across test stimuli provide an index of the extent to which listeners rely on F0 as a signal to speech category identity as a function of manipulations to the short-term input regularities across experimental conditions (acoustic, lexical, and acoustic + lexical) and blocks (canonical, reverse). Manipulations across conditions and blocks were not conveyed to participants, except inasmuch as response alternatives changed to match the stimuli.

Based on prior research, we predicted adaptive plasticity in reliance on F0 in the acoustic condition (Idemaru & Holt, 2011, 2014; Lehet & Holt, 2017; Schertz et al., 2016), but the influence of top-down lexical information was unknown. Therefore, to protect against the possibility of carryover effects should the experimental manipulations be effective in only some conditions, two groups of participants completed the experimental conditions in different orders. To foreshadow the results, the manipulation of lexical information had its intended effect and so data were collapsed across groups for all analyses and the group factor is not further examined. We next describe the detailed methods associated with each pretest and experimental condition.

2.3.2. Acoustic pretest

The acoustic pretest measured the baseline influence of F0 and VOT on /b/-/p/ categorization across a lexically neutral word-word (W-W) *beak-peak* stimulus space. On each

trial, listeners indicated whether they had heard *beak* or *peak* by pressing a key corresponding to orthographic *beak* and *peak* labels seen on the screen. Stimuli varied across a seven-step VOT series (sampled in 10-ms steps), paired with a high (F0 = 290 Hz) and a low (F0 = 230 Hz) F0 (see Fig. 1A). In all, there were 140 trials (2 F0 × 7 VOT × 10 repetitions) presented across about 6 min.

2.3.3. Lexical pretest

The lexical pretest assessed the influence of English word knowledge on /b/-/p/ categorization across lexically biased *beef-peef* (W-NW) and *peace-beace* (NW-W) acoustic spaces (Ganong, 1980). For both W-NW and NW-W contexts, participants categorized initial consonants as /b/ or /p/ across three perceptually ambiguous VOT values (5, 10, and 15 ms) at both high (F0 = 290 Hz) and low (F0 = 230 Hz) F0 (see Fig. 1B). On most trials (2 F0 × 3 VOT × 2 Lexical Contexts × 10 repetitions = 120 trials), participants saw two visual objects on the screen to indicate response options (a piece of meat to indicate *beef*, and a *peace* sign). These trials helped to reinforce the lexically biased context across the acoustically ambiguous stimuli. For a smaller proportion of trials (2 F0 × 1 VOT (10 ms) × 2 Lexical × 10 repetitions = 40 trials), participants saw *beef*, *peef*, *beace*, and *peace* as orthographic response options. These trials served as a test of the baseline influence of lexical context on categorization of the acoustically ambiguous speech input. In all, there were 160 trials presented across about 8 min.

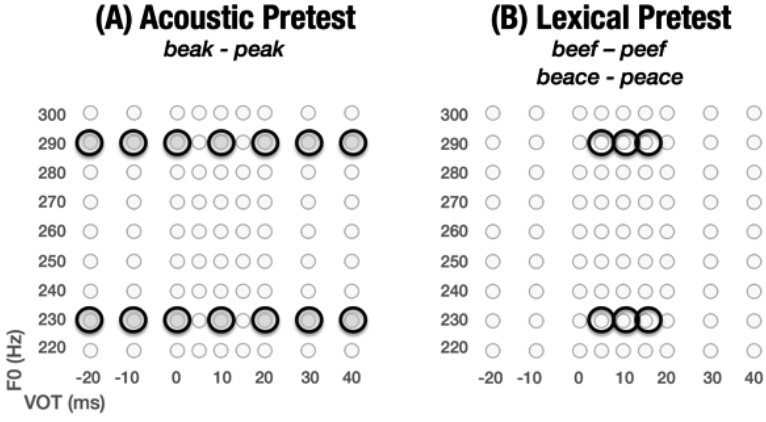

Fig. 1. Schematic representation of stimuli used in acoustic and lexical pretests. In each panel, the small symbols illustrate the full F0 × VOT stimulus space. The large symbols indicate stimuli presented in the experiment. (A) The acoustic pretest involved /b/-/p/ categorization of *beak-peak* (W-W) stimuli varying across seven VOT steps, at a high (F0 = 290 Hz) and low (F0 = 230 Hz) fundamental frequency, as shown by the large symbols. (B) The lexical pretest involved /b/-/p/ categorization across stimuli with three acoustically ambiguous VOT (5–15 ms) stimuli at a high (F0 = 290 Hz) and low (F0 = 230 Hz) F0, as shown by the large symbols. These stimuli were sampled across both *beef-peef* (W-NW) and *beace-peace* (NW-W) contexts to introduce a lexical bias toward /b/ and /p/, respectively, via the word frame.

2.3.4. Experimental conditions

Three additional conditions used the dimension-based statistical learning paradigm of prior research (Idemaru & Holt, 2011, 2014; Lehet & Holt, 2017; Liu & Holt, 2015; Schertz et al., 2016) to examine the core hypotheses (see Fig. 2). In this paradigm, the F0 × VOT correlation is manipulated to be consistent or inconsistent with typical English experience to track native-English listeners' weighting of acoustic dimensions. On *exposure trials* that comprise the majority of trials within a block (200 trials of 220 total trials, ~90%), the primary acoustic cue for /b/-/p/ categorization (Francis et al., 2008), VOT, unambiguously signals the speech category as /b/ or /p/. This presents the opportunity to manipulate the F0 × VOT correlation. In *canonical* blocks (Fig. 2A), F0 patterns with VOT in a manner that mirrors the long-term regularities of English such that long VOTs consistent with /p/ occur with high F0s and short VOTs consistent with /b/ occur with low F0s (Kingston & Diehl, 1994). In *reverse* blocks, an "artificial accent" is introduced that reverses the F0 × VOT correlation. Less frequent *test trials* for which stimuli have ambiguous VOT values and either a high or low F0 (see purple and orange symbols, Fig. 2A; 20 trials/block, ~10% of trials) are interspersed randomly throughout the exposure trials within both the canonical and reverse blocks. Test trials provide a means by which to assess how the short-term regularities of the exposure trials (canonical or reverse) affect perceptual reliance on F0 in /b/-/p/ categorization; since VOT is ambiguous (10 ms), only F0 (high = 290 Hz, low = 230 Hz) is available to signal /b/ versus /p/. Based on prior research, we hypothesize that category activation via the unambiguous acoustic VOT signal serves as a bottom-up, acoustic "teaching signal" to drive rapid adaptive plasticity in the extent to which the F0 of test trials is effective in signaling /b/-/p/ categories (Idemaru & Holt, 2011; Liu & Holt, 2015). In the present study, we include conditions that allow us to test whether phonetic category activation via top-down lexical knowledge may be a sufficient teaching signal when unambiguous bottom-up acoustic information (e.g., VOT) is unavailable. Across three conditions, the test trials are identical and are always presented in the lexically neutral *beak-peak* (W-W) context to support comparisons across conditions.

2.3.5. Acoustic condition

The acoustic condition modeled the approach of prior research (Idemaru & Holt, 2011, 2014; Lehet & Holt, 2017; Liu & Holt, 2015; Schertz et al., 2016; Zhang & Holt, 2018). Stimuli were sampled selectively across the *beak-peak* (W-W) stimulus space (see Fig. 2A). In this condition, there was no lexical bias to influence /b/-/p/ categorization. However, exposure trials were sampled such that acoustic, VOT information unambiguously signaled /b/-/p/ categories. Exposure stimuli with −20, −10 and 0 ms VOT reliably signaled /b/ whereas those with 20, 30, and 40 ms VOT reliably signaled /p/. In a first canonical block, VOT was paired with F0 in a manner that mirrored the typical correlation of these acoustic dimensions in English; lower F0s (220, 230, 240 Hz) were paired with VOTs signaling /b/ and higher F0s (280, 290, 300 Hz) were paired with VOTs signaling /p/. In a second reverse block, this relationship flipped so that the correlation of F0 and VOT was opposite that of English (see Fig. 2A). Across both canonical and reverse

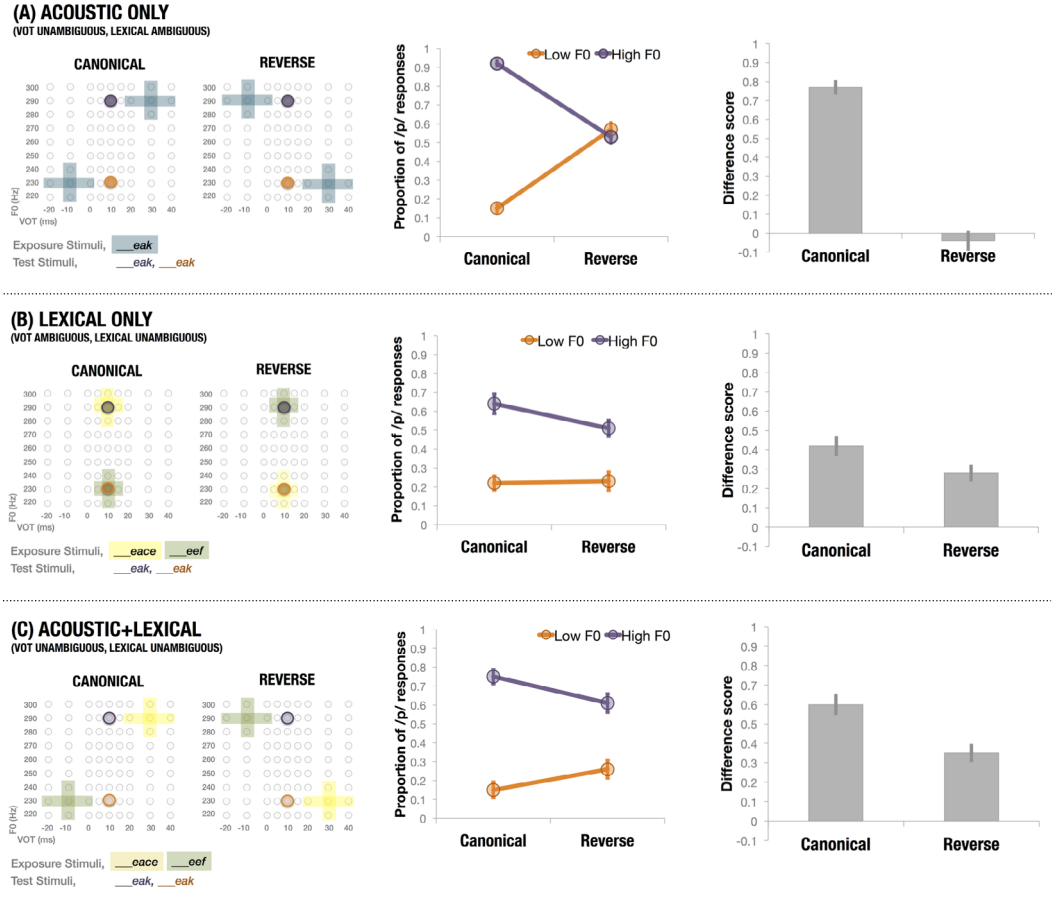

Fig. 2. Experiment conditions and data. The left panels illustrate the stimulus characteristics of the (A) acoustic only, (B) lexical only, and (C) acoustic + lexical conditions. For each condition, the unfilled dots illustrate stimuli sampling the full F0 × VOT stimulus space. Only a subset of stimuli were presented in each condition. The exposure stimuli are shown highlighted in color, with blue highlights corresponding to stimuli in *__eak* (W-W) context, yellow highlights to *__eace* (NW-W) context, and green highlights to *__eef* (W-NW) context. Test stimuli are shown as large filled circles with purple corresponding to high F0 (290 Hz) and orange to low F0 (230 Hz). Note that the test stimuli are identical across conditions, and are always presented in *__eak* (W-W) context. The middle and right panels show the data from each condition. The middle panels show average proportion of /p/ responses to the test stimuli (purple and orange filled circles in the left-most panels) with ambiguous VOT (10 ms) as a function of high (290 Hz) versus low (230 Hz) F0. The panels at the far right illustrate the same data as difference scores (proportion("p") responses for high F0 − low F0 test stimuli).

exposure trials, VOT unambiguously signaled /b/-/p/ categories; only the relationship of VOT to F0 varied across canonical and reverse blocks. Test trial categorization provided a measure of the extent to which experience with this short-term regularity affects reliance on F0 in /b/-/p/ categorization, which in prior studies has reliably been observed to rapidly change as a function of the regularities experienced across canonical versus

reverse blocks (e.g., Idemaru & Holt, 2011). Here, as in all experimental conditions, test trials were acoustically ambiguous VOT (10 ms) stimuli with high (F0 = 290 Hz) and low (F0 = 230 Hz) F0, presented in the lexically neutral *beak-peak* (W-W) context.

2.3.6. Lexical condition

There was also a lexical condition, as shown in Fig. 2B. In this condition, the exposure stimuli had perceptually ambiguous VOT (5, 10, 15 ms). Since VOT could not unambiguously signal /b/-/p/ categories, it was neutralized as cue to /b/-/p/ categorization. Instead, exposure stimuli were selectively sampled from *beef-peef* and *beace-peace* stimulus spaces (see Fig. 2B) such that lexical knowledge would support categorization of exposure stimuli as /b/ versus /p/ in a lexically consistent manner (i.e., /b/ for *beef-peef*, /p/ for *beace-peace*). Specifically, in the canonical block exposure stimuli were defined by *beace-peace* stimuli with ambiguous VOT and high F0s (280, 290, 300 Hz) and *beef-peef* stimuli with ambiguous VOT and low F0s (220, 230, 240 Hz). In a reverse block, exposure stimuli were defined such that *beef-peef* (with ambiguous VOT) had high F0 and *beace-peace* (with ambiguous VOT) had low F0. In this condition, we predicted that lexical knowledge of *beef* and *peace* would bias category-level activation to lexically consistent /b/ and /p/, respectively. To support this, the response options presented on screen for exposure trials were images corresponding to *beef* and *peace* (as in a portion of trials in the *lexical pretest*). Since the pairing of this lexical bias with F0 was such that it produced a canonical and a reverse short-term regularity, we predicted perceptual down-weighting of F0 akin to that observed via bottom-up acoustic F0 × VOT correlations in the acoustic only condition. We hypothesized that lexical information would evoke changes in reliance upon F0 in categorization of test stimuli via top-down selective activation of lexically consistent /b/ or /p/, as observed in the previous studies via bottom-up selective activation of /b/-/p/ via acoustic VOT information. The categorization of lexically neutral *beak-peak* (W-W) test trials with acoustically ambiguous VOT (10 ms) stimuli with high (F0 = 290 Hz) and low (F0 = 230 Hz) F0 provided the test of this hypothesis. For these trials, the response options on the screen were orthographic labels (*beak-peak*), as in the other experimental conditions.

2.3.7. Acoustic + lexical condition

There was also a condition with both acoustic and lexical information available to disambiguate speech input, as shown in Fig. 2C. In this condition, the exposure stimuli were sampled such that both acoustic (unambiguous VOT) and lexical information (top-down bias from *beef-peef* and *beace-peace* pairs) signaling /b/ versus /p/ were available in the input. In a canonical block, exposure trials were defined as perceptually unambiguous tokens with short VOT (consistent with /b/, −20, −10, 0 ms) presented in *beef-peef* context, with low F0 (220, 230, 240 Hz). Thus, perceptually unambiguous acoustic VOT input and English language knowledge of the word *beef* collaborate to signal /b/ paired with low F0, as typical in long-term English experience. Accordingly, unambiguous tokens with long VOT (consistent with /p/, 20, 30, 40 ms) were presented in *beace-peace* context, with high F0 (280, 290, 300 Hz). In the reverse bock, both the acoustic and

lexical information shifted to convey an F0 × VOT relationship opposite that typically experienced in English. Unambiguous short VOT tokens (consistent with /b/) were presented in *beef-peef* context (consistent with /b/) with a *high* F0 (typically correlated with /p/); unambiguous long VOT tokens were presented in *beace-peace* context with a *low* F0, contrary to long-term regularities of English (Fig. 2C). As in the other conditions, lexically neutral *beak-peak* (W-W) test trials with acoustically ambiguous VOT (10 ms) stimuli with high (F0 = 290 Hz) and low (F0 = 230 Hz) F0 served as the measure of the extent to which these short-term regularities impacted the effectiveness of F0 in signaling /b/ and /p/ speech categories.

The three experimental conditions differed only in the exposure trials (left panels, Fig. 2). As noted, test trials across conditions were identical; they possessed the same F0 and VOT (10 ms VOT; high F0 = 230 Hz, low F0 = 290 Hz) presented in *beak-peak* W-W pairs to eliminate lexical bias. Note that since all stimuli were created from the same base /bi/-/pi/ stimulus series, the underlying acoustics of exposure and test stimuli were identical for a particular point in the F0 × VOT acoustic space, except for the final consonant, across all conditions (i.e., *beace*, *beef*, *beak* have the same /bi/). Prior research indicates that the rapid adaptive plasticity with exposure stimuli generalizes robustly under these conditions (Liu & Holt, 2015). Nonetheless, note that manipulation of the lexical context resulted in heterogeneity in exposure stimuli. In the acoustic condition, both exposure and test stimuli were *beak-peak* tokens. In the lexical condition, the exposure involved *beef-peef* and *beace-peace* stimuli and test stimuli were *beak-peak* tokens. The acoustic + lexical condition was similar to the lexical condition, except that listeners heard tokens of *beef-peef* and *beace-peace* with unambiguous VOT.

3. Results

Data were analyzed using generalized linear mixed effects regression (GLMER) model (Breslow & Clayton, 1993) in R (lme4). The maximal random factor structure was modeled by including the categorical responses (i.e., voiced /b/ responses encoded as 0, and voiceless /p/ responses encoded as 1) as the dependent variable, and all possible factors justified by the experimental design as random factors (Barr, Levy, Scheepers, & Tily, 2013). The first model that converged included the by-subject and by-item intercepts only, and this model was selected as the base model. Fixed effects were assessed by testing the increase in model fit when each fixed factor was added to the base model. A likelihood ratio test was used to compare the fit between models (Baayen, Davidson, & Bates, 2008). The main effects of the fixed factors were assessed by adding each of the independent variables individually to the base model, and the interaction effects were assessed by comparing a model including these factors to a model including them and their interaction term (Chang, 2010; Mattys, Barden, & Samuel, 2014; Zhang & Samuel, 2015). All categorical factors were automatically coded by increasing numeric scales in R starting from 0. For example, when there were two levels within a factor, the level with

lower value was coded as 0 and the higher value was coded as 1. Factors with more than two levels used additional numbers to code for the additional levels.

3.1. Acoustic pretest

We first assessed the influence of F0 and VOT on /b/-/p/ categorization in the lexically neutral *beak-peak* context under baseline conditions with no short-term F0 × VOT correlation in the input. As shown in Fig. 3A, data were modeled as a 7 VOT × 2 F0 (high vs. low) design. There were main effects of both VOT [$\chi^2(6) = 35.93$, $p < .001$], and F0 [$\chi^2(1) = 12.13$, $p < .001$]. There also was an interaction between the two factors, $\chi^2(13) = 79.85$, $p < .001$. This is consistent with previous findings that the influence of F0 on voicing categorization is modulated by VOT, with the effect being the strongest when VOT is ambiguous (Kingston & Diehl, 1994; Kohler, 1982, 1984).

We next conducted a planned simple effect analysis on the stimuli with the most ambiguous VOT (10 ms) and high (290 Hz) versus low (230 Hz) F0 because test stimuli across the experimental blocks were defined by these acoustic characteristics (see Fig. 2). As shown in Fig. 3A, there was a robust effect of F0 on /b/-/p/ categorization when VOT was ambiguous, $\chi^2(1) = 9.42$, $p = .002$. Moreover, the directionality of this influence was in accord with the long-term covariation of F0 and VOT in English: *Beak-peak* stimuli with an ambiguous VOT were more often reported to be *peak* when F0 was higher ($M_{\text{HighF0}} = 0.85$, $SE = 0.04$, CI = [0.77, 0.93]) than when F0 was lower ($M_{\text{LowF0}} = 0.32$, $SE = 0.06$, CI = [0.20, 0.44]). In this baseline block in which there was no short-term information of an F0 × VOT correlation, /b/-/p/ speech categorization reflected long-term regularities of English. Both VOT and F0 affected assessments of category membership.

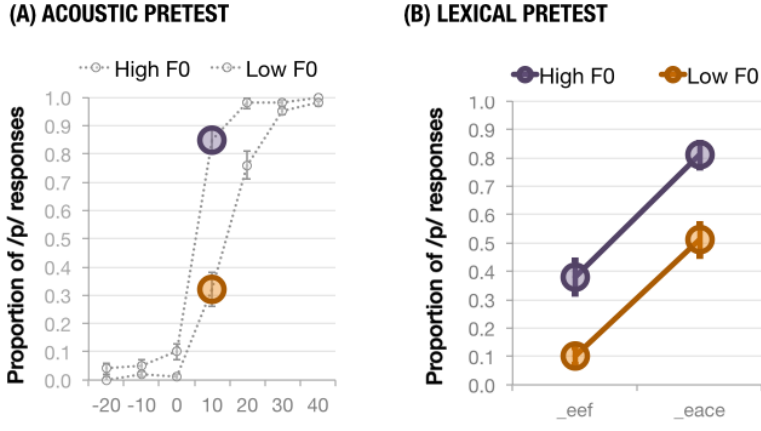

Fig. 3. Results of acoustic and lexical pretests. The stimulus F0 affected /b/-/p/ categorization when VOT was the most ambiguous (VOT = 10 ms). (A) Acoustic pretest. This was evident in the acoustic pretest for which there was no lexical bias (*beak-peak*). (B) Lexical pretest. In the lexical pretest, both lexical context (*_eef*, *_eace*) and acoustic F0 (high, low) influenced /b/-/p/ categorization. Error bars are standard error of the mean.

3.2. Lexical pretest

We next assessed the influence of English word knowledge on /b/-/p/ categorization across lexically biased *beef-peef* (W-NW) and *peace-beace* (NW-W) contexts, for the trials with ambiguous VOT (10 ms) and orthographic response labels that did not reinforce lexical interpretation of the stimuli. As shown in Fig. 3B, data were modeled as a Lexical Context (*_eef* vs. *_eace*) × F0 (high vs. low) design. There was a main effect of lexical context, $\chi^2(1) = 28.21$, $p < .001$, for the acoustically ambiguous VOT stimulus that serves as the test stimulus in the experimental conditions. Participants categorized these stimuli more often as /p/ in *__eace* context ($M_{\_eace} = 0.66$, $SE = 0.05$, CI = [0.56, 0.75]) than in *__eef* context ($M_{\_eef} = 0.24$, $SE = 0.04$, CI = [0.17, 0.32]). There was also a main effect of F0, $\chi^2(1) = 32.45$, $p < .0016$, indicating that participants were more likely to categorize the ambiguous 10-ms VOT sound as /p/ when the F0 was high ($M_{\text{HighF0}} = 0.60$, $SE = 0.04$, CI = [0.51, 0.69]) than when F0 was low ($M_{\text{LowF0}} = 0.31$, $SE = 0.04$, CI = [0.23, 0.38]). There was no interaction, $\chi^2(3) = 0.33$, $p = .563$.

In all, the pretest results confirm that when VOT is acoustically ambiguous /b/-/p/ categorization is affected by both lexical context and acoustic F0 information within the stimulus sets created for the present experiment, as expected from prior research (Ganong, 1980; Idemaru & Holt, 2011; Kingston & Diehl, 1994).

3.3. Exposure trials across experimental conditions

We examined categorization across exposure trials, which comprised the majority (90%) of trials in the experimental blocks. These trials involved putatively perceptually *unambiguous* information with which to resolve /b/-/p/ categorization, via either bottom-up acoustic information (acoustic condition) or top-down lexical knowledge (lexical condition) or both (acoustic + lexical condition) and conveyed short-term regularities consistent with English (canonical blocks) or inconsistent with English (reverse blocks, "artificial accent") for each of the three experimental conditions. The results confirm that listeners were able to resolve the /b/-/p/ categories with high accuracy across exposure trials, as shown in Table 2.

These results assure us that the exposure stimuli served the intended role of pushing /b/-/p/ categorization toward one phonetic category alternative or the other, as a function of either bottom-up acoustic information, top-down lexical knowledge, or their combination. Listeners made use of VOT, F0, and lexical context in informing /b/-/p/ category decisions.

3.4. Test trials across experimental conditions

Following the approach of prior research (Idemaru & Holt, 2011, 2014; Lehet & Holt, 2017; Liu & Holt, 2015; Schertz et al., 2016), we analyzed the test trials to test our core hypotheses. The logic of this approach is that since the test trials had acoustically ambiguous VOT in a lexically neutral *beak-peak* context, only F0 provided information to inform /b/-/p/ categorization. Thus, analysis of test trials provides a means of tracking

Table 2
Proportion /p/ responses to exposure trials across conditions. The mean proportion of /p/ responses as a function of condition and block for exposure trials demonstrate that exposure trials were perceptually unambiguous as intended. The standard error of the mean is shown in parentheses

		Shorter VOT	Longer VOT	__*eef*	__*eace*	Shorter VOT and __*eef*	Longer VOT and __*eace*
Acoustic	Canonical	0.03 (0.02)	0.98 (0.02)				
	Reverse	0.11 (0.04)	0.93 (0.03)				
Lexical	Canonical			0.09 (0.02)	0.96 (0.02)		
	Reverse			0.21 (0.24)	0.95 (0.01)		
Acoustic + Lexical	Canonical					0.02 (0.02)	0.99 (0.01)
	Reverse					0.04 (0.02)	0.98 (0.01)

the effectiveness of F0 in signaling /b/-/p/ categories as a function of the short-term regularities experienced in the input across exposure trials in canonical versus reverse blocks over experimental conditions.

Categorization responses for the test stimuli were modeled as a condition (acoustic, lexical, acoustic + lexical) × block (canonical, reverse) × F0 (high, low) design. The analysis revealed main effects of condition [$\chi^2(2) = 53.54$, $p < .001$], and F0 [$\chi^2(1) = 11.03$, $p < .001$], and no effect of block [$\chi^2(1) = 1.36$, $p = .243$]. A block × F0 interaction revealed robust modulation of the effectiveness of F0 in /b/-/p/ categorization as a function of short-term experience, $\chi^2(3) = 172.61$, $p < .001$. F0 was more effective in signaling /b/-/p/ category membership in the canonical blocks, $z = 22.16$, $p < .001$, than in the reverse blocks, $z = 8.27$, $p < .001$. There was an interaction of condition × block, $\chi^2(2) = 59.01$, $p < .001$. There was a simple effect of block only in the lexical condition, $z = 3.63$, $p < .001$; ps > .400 in other two conditions. There was no interaction between condition and F0, $\chi^2(5) = 7.20$, $p = .202$. Finally, there was a three-way interaction across condition, block, and test stimulus F0, $\chi^2(3) = 32.46$, $p < .001$, indicating that the block × F0 interaction was modulated by condition. We next describe these patterns in detail as a function of the experimental conditions.

3.4.1. Acoustic condition

The middle panel of Fig. 2A plots the average proportion /p/ responses for high and low F0 test stimuli as a function of canonical and reverse blocks for the acoustic condition. There was a main effect of F0, indicating that listeners relied on F0 to make /b/-/p/ categorization decisions when VOT was ambiguous [$\chi^2(1) = 9.91$, $p = .002$]. There was no main effect of block, $\chi^2(1) = 0.37$, $p = .543$, indicating no overall shift in average /p/ responses as a function of block. Most importantly, the block × F0 interaction revealed that the relationship of F0 and VOT experienced across exposure stimuli within the canonical and reverse blocks impacted the influence of F0 in /b/-/p/ categorization [$\chi^2(3) = 223.75$, $p < .001$]. The simple effect of F0 was robust in the canonical block ($z = 14.66$, $p < .001$; $M_{\text{HighF0}} = 0.85$, $SE = 0.04$, CI = [0.77, 0.93]; $M_{\text{LowF0}} = 0.15$,

SE = 0.03, CI = [0.09, 0.21]), but not in the reverse block (z = 0.063, p = .528 (M_{HighF0} = 0.53, SE = 0.04, CI = [0.45, 0.62]; M_{LowF0} = 0.57, SE = 0.04, CI = [0.48, 0.66]), replicating prior literature. When short-term regularities in speech input departed from long-term regularities of English, listeners rapidly down-weighted reliance on F0 (Idemaru & Holt, 2011, 2014; Lehet & Holt, 2017; Schertz et al., 2016). The rightmost panel of Fig. 2A shows these same data as difference scores in the proportion of /p/ responses to high versus low F0 test stimuli as a function of short-term experience in the canonical and reverse block.

3.4.2. Lexical condition

The middle panel of Fig. 2B shows the average /p/ categorization responses for high and low F0 test stimuli as a function of block for the lexical condition, with the rightmost panel plotting the same data as difference scores across test stimuli. There was a main effect of block [$\chi^2(1)$ = 5.12, p = .024] and F0 [$\chi^2(1)$ = 8.76, p = .003], and an interaction [$\chi^2(3)$ = 20.40, p < .001]. The simple effect of F0 was present in both the canonical (z = 9.92, p < .001; M_{HighF0} = 0.64, SE = 0.05, CI = [0.54, 0.75]; M_{LowF0} = 0.22, SE = 0.04, CI = [0.14, 0.30]) and reverse blocks (z = 2.55, p = .011; M_{HighF0} = 0.51, SE = 0.04, CI = [0.42, 0.60]; M_{LowF0} = 0.24, SE = 0.05, CI = [0.14, 0.33]). This pattern indicates that listeners continued to rely upon F0 in /b/-/p/ categorization in the reverse block, but (as indicated by the interaction) the influence of F0 was diminished in the reverse block compared to the canonical block. Lexical context, in the absence of acoustically unambiguous bottom-up information to differentiate /b/ and /p/ categories across exposure stimuli, was sufficient to evoke down-weighting of the effectiveness of F0 in signaling /b/-/p/ categories. As is visually apparent in the difference scores plotted in Fig. 2A,B, the extent of down-weighting was weaker in the lexical only condition than in the acoustic only condition, but nonetheless present; this is evident in the effect sizes for the block × F0 interactions across conditions, as well. Top-down lexical knowledge appears to have reliably driven adaptive plasticity in the effectiveness of F0 in signaling /b/-/p/ categorization, albeit somewhat less effectively than perceptually unambiguous bottom-up acoustic information. We return to consider this in Section 4.

3.4.3. Acoustic + lexical condition

The middle panel of Fig. 2C shows the average /p/ categorization responses for high and low F0 test stimuli as a function of block for the acoustic + lexical condition. There was a main effect of F0 [$\chi^2(1)$ = 10.05, p = .001], and there was no effect of block, [$\chi^2(1)$ = 0.28, p = .60]. Of most importance to our hypotheses, there was a block × F0 interaction, indicating a decrease in the diagnosticity of F0 for /b/-/p/ categorization in the reverse compared to the canonical block [$\chi^2(3)$ = 34.44, p < .001]. There was a simple effect of F0 in both the Canonical (z = 12.96, p < .001; M_{HighF0} = 0.75, SE = 0.04, CI = [0.67, 0.83]; M_{LowF0} = 0.15, SE = 0.04, CI = [0.06, 0.23]) and reverse blocks (z = 4.83, p < .001; M_{HighF0} = 0.61, SE = 0.05, CI = [0.51, 0.71]; M_{LowF0} = 0.26, SE = 0.05, CI = [0.16, 0.35]), but listeners relied on F0 less in the reverse, compared to the canonical block. Thus, down-weighting of F0 as a function of short-term experience that

departs from the norm is observed when both lexical and acoustic information are available in the signal to inform categorization of exposure stimuli. Interestingly, although the acoustic information available across blocks was identical to that available in the acoustic condition, the extent of down-weighting observed was somewhat weaker in the acoustic + lexical condition. This is apparent visually in Fig. 2A versus Fig. 2C, in the block × F0 interaction effect size across conditions, and the presence of the three-way interaction in the omnibus analysis, indicating a modulation of the degree of down-weighting across conditions.

3.5. General discussion

Speech communication presents an excellent testbed for investigating the stability-plasticity dilemma faced by all cognitive systems. On the one hand, there is pressure to align with long-term input regularities to guide behavior effectively and efficiently. In speech, this is achieved in part through acquisition of robust native-language speech categories that reflect the nuanced relationships of the multiple sensory dimensions associated with speech categories (Holt & Lotto, 2006; Idemaru et al., 2012; McMurray & Jongman, 2011; Toscano & McMurray, 2010). On the other hand, there is pressure to flexibly adapt to short-term input that deviates from these long-term norms. In speech, this can involve adapting to regularities associated with a distinct dialect or accent, or a conversation partner who is suffering from a head cold. Speech communication often takes place across input that is an imperfect match to the long-term speech regularities that have shaped the mapping of speech acoustics to linguistically significant representations like phonemes and words. Despite the challenges that these short-term deviations from typical regularities introduce, speech perception rapidly adapts. Across multiple empirical paradigms, this rapid adaptive plasticity has been shown to be supported by the presence of an information source that disambiguates the short-term input acoustics. Diverse information sources can contribute, including acoustic (e.g., Idemaru & Holt, 2011), visual (e.g., Vroomen et al., 2007), or lexical (e.g., Norris et al., 2003) information that disambiguates speech input and leads to perceptual shifts that endure even when that information is no longer available. But these effects have been investigated independently and often in somewhat different paradigms.

The present study created a context in which disambiguating lexical and acoustic information sources could be jointly examined, with the aim of testing the hypothesis that resolution of phonetic category activation, whether through bottom-up acoustic information or top-down lexical information, drives adaptive plasticity in speech perception.

The acoustic and lexical pretests established that both acoustic and lexical information influenced recognition of acoustically ambiguous speech, disambiguating perception in a manner consistent with the long-term regularities of English experience. Consistent with the co-variation of F0 and VOT in English-language experience, native-English adults were more likely to categorize a sound with an acoustically ambiguous VOT as /b/ when F0 was low. The same sound was more often categorized as /p/ when F0 was high. Likewise, lexical knowledge also influenced speech categorization. In lexically biased contexts, listeners were more likely to categorize speech with acoustically ambiguous VOT

as lexically consistent (/b/ in the context of __eef and /p/ in the context of __eace). When bottom-up VOT information is perceptually ambiguous, listeners rely on both bottom-up acoustic information about F0 and word knowledge to resolve the speech input as /b/ versus /p/. In all, the pretest results set the stage for us to ask whether phonetic category resolution through top-down lexical information is sufficient to drive rapid adaptive plasticity in the extent to which listeners rely on F0 to inform /b/-/p/ categorization.

The question under investigation is whether adaptive re-weighting of F0 as a cue to /b/-/p/ categorization is driven by activation of the /b/ versus /p/ categories via unambiguous input, whether the input is acoustic or lexical. Replicating prior research, the disambiguating acoustic information from VOT was sufficient to drive adaptive plasticity of weighting F0 (Idemaru & Holt, 2011, 2014; Lehet & Holt, 2017; Liu & Holt, 2015; Schertz et al., 2016; Zhang & Holt, 2018). Listeners quickly re-weighted reliance on F0 in speech categorization in the reverse block, within which the correlation between F0 × VOT dimensions was opposite that of long-term English experience. Categorization of exposure trials with unambiguous VOT was highly selective even in the context of the artificial accent that reversed the relationship of VOT to F0 in speech input. This is consistent with the possibility that selective phonetic category activation via bottom-up unambiguous VOT information plays a role in the down-weighting of F0 observed across test trials.

Crucially, if selective phonetic category activation drives this adaptive re-weighting, then it should persist even when bottom-up VOT information is rendered ambiguous but top-down information about word knowledge is available to support selective phonetic category activation. In this way, the lexical condition stimuli were organized such that there was no bottom-up acoustic information from VOT to resolve phonetic category membership. However, lexical knowledge was available to resolve phonetic categories as /b/ or /p/ in a manner biased toward real English words. Moreover, the lexically resolved phonetic categories were paired with high versus low F0 in such a way as to convey the F0 × VOT relationship typical of English, or the reversed relationship of the "artificial accent." Even without bottom-up acoustic information to drive perceptual tuning, the listeners relied less on F0 in /b/-/p/ categorization in the context of the artificial accent. Thus, top-down lexical knowledge appears to be sufficient to tune the perceptual weighting of acoustic dimensions in speech categorization. This is consistent with selective phonetic category activation, biased by either bottom-up or top-down information to disambiguate category identity, driving adaptive plasticity effects on the dynamic reweighting observed for how effectively specific acoustic input dimensions contribute to speech recognition. An important implication of this pattern of adaptive plasticity is that the very dimensions that define perceptual categories are dynamically, and rapidly, adjusted in online speech processing to accommodate regularities in the ambient speech environment. The manner by which acoustic dimensions map to speech categories and words is not rigidly fixed by long-term experience. Rather, the "feature space" serving speech recognition flexibly, and rather rapidly, adapts to local regularities.

In this regard, it is important to be clear that there was one source of acoustic information available to convey /b/ versus /p/ category membership of exposure stimuli in the lexical condition: F0. Stimuli had either high or low F0, which is a secondary cue to /b/-/p/

category membership. As is evident from the acoustic pretest data, bottom-up acoustic F0 information can be sufficient to inform phonetic category membership. It is reasonable to question, then, whether it is possible that this bottom-up acoustic F0 information—rather than lexical information—could have been responsible for the re-weighting observed in the lexical condition. However, the directionality of F0 re-weighting observed across canonical and reverse lexical blocks indicates it did not. The perceptually ambiguous VOT made it an unreliable signal of /b/ versus /p/ category in the lexical condition. Yet, although F0 was present to potentially signal /b/ when it was low and /p/ when it was high (consistent with long-term norms), it was constant across the canonical and reverse blocks' exposure trials. In this way, judged by the F0 input alone, there was no "artificial accent" or difference in the short-term input regularities across canonical and reverse blocks. Thus, the observation of re-weighting in the reverse block in which F0 was mismatched with the lexically resolved phonetic categories relative to English experience indicates that it was the lexical, not the acoustic F0, information responsible for the pattern of perceptual results. Re-weighting of F0 therefore must arise from the pairing of F0 with the phonetic category implied by lexical context, and the fact that it mismatched long-term English regularities in the reverse block.

It is important to note that the re-weighting of the diagnosticity of F0 in /b/-/p/ categorization was not lexically specific. The lexical condition was structured such that listeners experienced exposure trials conveying the short-term regularity via lexically biased beef-peef and beace-peace word/nonword frames. Re-weighting was measured across lexically neutral beak-peak word-word test trials equivalent to those examined in the other experimental conditions. Thus, the F0 × VOT correlation implied by lexical activation of /b/-/p/ categories across short-term exposure exerted an influence on speech categorization that was not limited to identical lexical contexts. This finding is consistent with prior research demonstrating generalization of acoustically driven perceptual tuning from nonlexical to lexical items (Lehet & Holt, 2020; Liu & Holt, 2015). The present results extend this pattern of generalization across even more distinct contexts. Importantly, these observations are supported by the results of the acoustic + lexical condition, which moved toward the goal of integrating multiple disambiguating information sources in examinations of adaptive plasticity in speech perception.

Interestingly, the degree of re-weighting of F0 observed from the canonical to the reverse blocks was dampened in both the lexical and the acoustic + lexical conditions relative to the acoustic condition. It is somewhat tempting to suggest that top-down signals may be less effective at driving adaptive plasticity than bottom-up signals that resolve acoustic speech ambiguity and selectively activate phonetic category representations. However, perhaps more likely, this may instead relate to less-than-complete generalization of re-weighting (to test stimuli, beak-peak) from the contexts in which the artificial accent was experienced (beef-peef, beace-peace).

In the present study, we prioritized having identical test stimuli across conditions (beak-peak) and including an acoustic condition aligned with prior demonstrations in the literature. As a result, adaptive plasticity in the acoustic condition was observed across tokens with a context common to that experienced in the artificial accent (beak-peak),

whereas perceptual tuning in the lexical and lexical + acoustic conditions necessarily required generalization from experience with the artificial accent across beef-peef and beace-peace exposure stimuli to beak-peak test stimuli. This put the greatest generalization demands on the lexical and acoustic + lexical conditions, allowing us to make a highly conservative test of the prediction that top-down lexical knowledge can drive perceptual re-weighting via phonetic category activation. Nonetheless, this aspect of the experiment is important to consider, and cross-condition comparisons of the magnitude of re-weighting should be made cautiously. Future studies integrating a greater diversity of lexical and non-lexical generalization contexts among test trials will help to establish the impact of generalization on the results, an important open issue across all forms of adaptive plasticity (Dahan & Mead, 2010; Eisner & McQueen, 2005; Idemaru & Holt, 2014; Kraljic & Samuel, 2006, 2007; Lehet & Holt, 2020; Liu & Holt, 2015; Reinisch & Holt, 2014; Reinisch, Wozny, Mitterer, & Holt, 2014). Here, we opted for consistent test tokens across conditions to conservatively test the central hypothesis that phonetic category activation drives this form of adaptive plasticity.

As an aside, we note that the artificial accent presented in the current study is a rather major shift in short-term speech input statistics; the correlation between two robust cues to speech categorization is reversed. The artificial nature of this "accent" provides both a well-controlled testbed for examining the statistical regularities of short-term input that produce adaptive plasticity, and a test of the system's ability to adapt robustly. A reversal in the correlation between two acoustic input dimensions is a strong shift in speech input regularities, but it has precedent in natural languages. For example, native-English speakers learning the Korean three-way stop consonant distinction can fail to produce the correct Korean F0 × VOT relationship (Kim & Lotto, 2002). Nonnative English spoken by Japanese speakers often reverses the relationship of the second and third formant frequencies that typify native English speech (Lotto et al., 2004). Moreover, although Scottish English /i/-/I/ distinction differs almost exclusively in spectral information, speakers from the South of England produce the same vowels with a considerable durational difference and a less substantial spectral difference (Escudero, 2001). The present approach allows us to control short-term speech input regularities to investigate the flexibility of perception while still using stimulus materials that are highly natural.

In a larger context, the present data contribute to understanding the mechanistic bases of adaptive plasticity in speech perception. To date, modeling of adaptive plasticity effects in speech (Kleinschmidt & Jaeger, 2015) has focused on a computational level of analysis (Marr, 1982) that describes what the system does and why it does these things. For example, Kleinschmidt and Jaeger (2015) conceptualized the computational demands of adaptive plasticity in speech as a belief updating process whereby listeners accumulate speech input statistics within a listening environment and adapt to these local statistical regularities. Nonetheless, considered from Marr's framework of theory development (1982), we presently have a poor algorithmic understanding of how these computational demands are met. In the present research, the central hypothesis that phonetic category activation, whether through top-down or bottom-up information, drives the perceptual re-weighting of F0 is directed at the algorithmic level regarding specific mechanistic questions of how the system

does what it does and which representations and processes are involved (Marr, 1982). Though it remains for future work to propose a detailed algorithmic model, the present results argue for a role for phonetic category activation as a driver of adaptive plasticity.

In this way, the present results are consistent with a working model put forward by Guediche et al. (2014), which suggests adaptive plasticity in speech can be considered as a form of sensori-cognitive adaptation whereby long-term speech representations activated by unambiguous elements of the input (lexical context, a dominant acoustic dimension) provide predictions about the sensory input based on the past experience that gave rise to the cognitive representation. For example, the activation of a phonetic category /b/ by an unambiguous short VOT would provide a prediction for low-frequency F0, by virtue of the manner in which the system has organized to support long-term regularities of English in which short VOTs are associated with lower frequency F0.

Upon encountering input that mismatches these predictions (e.g., an artificial accent that reverses the correlation of dimensions), an error signal may be generated to support rapid adjustment to minimize future error. Guediche et al. (2014) discuss evidence for a neurologically plausible conceptual model of adaptive plasticity, potentially dependent on the cerebellum, as the engine for error-driven supervised learning to drive adaptive adjustments, analogous to mechanisms in sensorimotor adaptation. According to this model, adaptive plasticity hinges on the activation of a cognitive representation to generate the prediction that drives perceptual tuning. The present data provide support for the key prediction that phonetic category activation, whether via bottom-up acoustic information or top-down lexical information, is sufficient to elicit adaptive plasticity that modulates the effectiveness of an acoustic dimension in signaling speech categories as a function of short-term input regularities that deviate from the norm.

Broadening this model, Liu and Holt (2015) proposed that re-weighting of the diagnosticity of an acoustic dimension in response to exposure to an artificial accent with dimension regularities that differ from long-term experience may be accounted for by a multilevel interactive representational network with assumptions similar to speech recognition models like TRACE (McClelland & Elman, 1986; Mirman et al., 2006). In this conceptualization, the initial connection weights among representations are related to the perceptual weights learned through long-term regularities in speech input. In this way, the baseline reliance on F0, VOT, and lexical context measured by the present acoustic and lexical pretests can be understood to approximate the relative strength of initial connection weights in the network. To accommodate the rapid adaptive plasticity observed in the present results, these weights would need to be modifiable. Prior modeling efforts have incorporated Hebbian learning to adjust connection weights to account for lexically driven perceptual tuning (Hebb-TRACE; Mirman et al., 2006). Liu and Holt (2015) noted that although this approach could capture patterns of acoustically driven adaptive plasticity like those observed in the present acoustic condition, strictly Hebbian learning may be too sluggish to account for the rapidity of dimensional reweighting observed here and in prior studies (Idemaru & Holt, 2011, 2014; Liu & Holt, 2015; see Guediche et al., 2014 for discussion). Guediche et al. (2014) proposed that supervised learning mechanisms may be better aligned with the rapidity of adaptive plasticity because the internal model of a target representation (e.g., the established

connection weights) serves as an internal prediction of expected stimulus qualities associated with a particular representation (e.g., a phonetic category). When the stimulus input deviates from these expectations as in the case of an accent, error-based supervised learning can rapidly adjust the representation or the connection weights. In this context, perceptually unambiguous bottom-up VOT information (as in the acoustic condition) can serve as a "teaching signal" that is sufficient to robustly activate /b/-/p/ categories based on strong connection weights established by long-term experience. Owing to the representation of long-term distributional regularities of how acoustic dimensions map to phonetic categories, this activation provides predictions about how other (e.g., F0) acoustic dimensions typically map to the category. These predictions may be compared with the actual sensory input, with discrepancies resulting in an internally generated error signal that can drive adaptive adjustments of the internal prediction to improve alignment of future predictions with incoming input, as a biologically plausible mechanism widely attested in cognitive and motor systems (Doya, 2000; Wolpert, Diedrichsen, & Flanagan, 2011). The present results support the hypothesis that adaptive plasticity of the mapping of acoustic dimensions to speech representations is driven by phonetic-category-level activation; robust activation of a phonetic category, whether from bottom-up or top-down information, is sufficient to drive rapid adjustments in how incoming acoustic input maps to speech representations.

In this regard, the present results and their relationship to a sensori-cognitive adaptation characterization of adaptive plasticity also speak to a long-debated issue in models of spoken word recognition—whether the nature of information processing is feedforward or interactive. Feedforward models (e.g., Norris et al., 2000) posit that the flow of information in the perceptual system is strictly bottom-up. In these models, "top-down" effects of lexical knowledge on phonetic categorization emerge as a result of integration of phonetic and lexical information at a later decision stage. In contrast, interactive models allow "top-down" lexical knowledge to directly modulate activation of pre-lexical representations. Although the models have very distinct architectures, it has been very difficult to empirically distinguish them in practice (see McClelland et al., 2006). This is especially the case since lexically driven adaptive plasticity has been accommodated in feedforward models by feedback for learning, based on the hypothesis that the system is feedforward for online perception, with interactive feedback only for learning (Norris et al., 2003). The present results are particularly interesting with regard to this theoretical divide because they provide fine-grained evidence that top-down lexical information can influence the effectiveness of a particular acoustic dimension in its ability to signal category identity. This top-down influence thus reaches very early levels of speech representation, and it must be accommodated by any architecture modeling speech perception.

References

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412. https://doi.org/10.1016/j.jml.2007.12.005

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Boersma, P., & Weenink, D. (2017). Praat: doing phonetics by computer [Computer program]. Version 6.1.38. Retrieved from http://www.praat.org/ Accessed January 2, 2021.

Breslow, N. E., & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, *88*(421), 9–25.

Chang, L. (2010). Using lme4. University of Arizona. Retrieved from www.u.arizona.edu/~ljchang/NewSite/papers/LME4_HO.pdf Accessed December 12, 2013.

Dahan, D., & Mead, R. L. (2010). Context-conditioned generalization in adaptation to distorted speech. *Journal of Experimental Psychology: Human Perception and Performance*, *36*(3), 704.

Doya, K. (2000). Complementary roles of basal ganglia and cerebellum in learning and motor control. *Current Opinion in Neurobiology*, *10*(6), 732–739.

Eisner, F., & McQueen, J. M. (2005). The specificity of perceptual learning in speech processing. *Perception & Psychophysics*, *67*(2), 224–238.

Escudero, P. (2001). The role of the input in the development of L1 and L2 sound contrasts: Language-specific cue weighting for vowels. In A. H. L. Do L. Domínguez & A. Johansen (Eds.), *Proceedings of the 25th annual Boston University conference on language development*, Somerville, MA: Cascadilla Press.

Francis, A., 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), 1234–1251.

Ganong, W. F. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology. Human Perception and Performance*, *6*(1), 110–125.

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

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 Pt 1), 3059–3071. https://doi.org/10.1121/1.2188377

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

Idemaru, K., & Holt, L. L. (2014). Specificity of dimension-based statistical learning in word recognition. *Journal of Experimental Psychology. Human Perception and Performance*, *40*(3), 1009–1021. https://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), 3950–3964. https://doi.org/10.1121/1.4765076

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.

Kim, M., & Lotto, A. (2002). An investigation of acoustic characteristics of Korean stops produced by non-heritage learners. *The Korean Language in America*, *7*, 177–187. Retrieved from http://www.jstor.org/stable/42922194 Accessed February 2, 2021.

Kingston, J., & Diehl, R. L. (1994). Phonetic knowledge. *Language*, *70*, 419–454.

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

Kohler, K. J. (1982). F0 in the production of lenis and fortis plosives. *Phonetica*, *39*(4–5), 199–218.

Kohler, K. J. (1984). Phonetic explanation in phonology: The feature fortis/lenis. *Phonetica*, *41*(3), 150–174.

Kondaurova, M. V., & Francis, A. L. (2008). The relationship between native allophonic experience with vowel duration and perception of the English tense/lax vowel contrast by Spanish and Russian listeners. *The Journal of the Acoustical Society of America*, *124*(6), 3959. https://doi.org/10.1121/1.2999341

Kraljic, T., & Samuel, A. G. (2006). Generalization in perceptual learning for speech. *Psychonomic Bulletin & Review*, *13*(2), 262–268.

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

Lehet, M., & Holt, L. (2017). Dimension-based statistical learning affects both speech perception and production. *Cognitive Science*, *41*(Suppl. 4), 885–912. https://doi.org/10.1111/cogs.12413

Lehet, M., & Holt, L. L. (2020). Nevertheless, it persists: Dimension-based statistical learning and normalization of speech impact different levels of perceptual processing. *Cognition*, *202*, 104328. https://doi.org/10.1016/j.cognition.2020.104328

Liu, R., & Holt, L. L. (2015). Dimension-based statistical learning of vowels. *Journal of Experimental Psychology. Human Perception and Performance*, *41*(6), 1783–1798. https://doi.org/10.1037/xhp0000092

Lotto, A., Sato, M., & Diehl, R. (2004). Mapping the task for the second language learner: The case of Japanese acquisition of /r/ and /l/. In J. Slifka (Ed.). *From sound to sense: 50+ years of discoveries in speech communication*, Cambridge, MA: Research Laboratory of Electronics at MIT.

Marr, D. (1982). *Vision. A computational investigation into the human representation and processing of visual information*. Cambridge, MA: MIT Press.

Mattys, S. L., Barden, K., & Samuel, A. G. (2014). Extrinsic cognitive load impairs low-level speech perception. *Psychonomic Bulletin & Review*, *21*(3), 748–754. https://doi.org/10.3758/s13423-013-0544-7

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

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, *18*, 1–86. https://doi.org/10.1016/0010-0285(86)90015-0

McClelland, J. L., Mirman, D., & Holt, L. L. (2006). Are there interactive processes in speech perception? *Trends in Cognitive Sciences*, *10*(8), 363–369. https://doi.org/10.1016/j.tics.2006.06.007

McCloskey, M., & Cohen, N. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. In G. H. Bower (Ed.), *The Psychology of Learning and Motivation*, *24*, 109–164.

McMurray, B., & Jongman, A. (2011). What information is necessary for speech categorization? Harnessing variability in the speech signal by integrating cues computed relative to expectations. *Psychological Review*, *118*(2), 219.

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. https://doi.org/10.3758/BF03213909

Norris, D., McQueen, J. M., & Cutler, A. (2000). Merging information in speech recognition: Feedback is never necessary. *Behavior and Brain Science*, *23*(3), 299–325. https://doi.org/10.1017/s0140525x00003241

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology*, *47*(2), 204–238. https://doi.org/10.1016/S0010-0285(03)00006-9

Ratcliff, R. (1990). Connectionist models of recognition memory: Constraints imposed by learning and forgetting functions. *Psychological Review*, *97*, 285–308.

Reinisch, E., & Holt, L. L. (2014). Lexically guided phonetic retuning of foreign-accented speech and its generalization. *J Exp Psychol Hum Percept Perform*, *40*(2), 539–555. https://doi.org/10.1037/a0034409

Reinisch, E., Wozny, D. R., Mitterer, H., & Holt, L. L. (2014). Phonetic category recalibration: What are the categories?. *J Phon*, *45*, 91–105. https://doi.org/10.1016/j.wocn.2014.04.002

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

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

Schwab, E. C., Nusbaum, H. C., & Pisoni, D. B. (1985). Some effects of training on the perception of synthetic speech. *Human Factors*, *27*(4), 395–408. https://doi.org/10.1177/001872088502700404

Toscano, J. C., & McMurray, B. (2010). Cue integration with categories: Weighting acoustic cues in speech using unsupervised learning and distributional statistics. *Cognitive Science*, *34*(3), 434–464. https://doi.org/10.1111/j.1551-6709.2009.01077.x

Vroomen, J., van Linden, S., de Gelder, B., & Bertelson, P. (2007). Visual recalibration and selective adaptation in auditory–visual speech perception: Contrasting build-up courses. *Neuropsychologia*, *45*(3), 572–577. https://doi.org/10.1016/j.neuropsychologia.2006.01.031

Wolpert, D. M., Diedrichsen, J., & Flanagan, J. R. (2011). Principles of sensorimotor learning. *Nature Reviews Neuroscience*, *12*(12), 739–751. https://doi.org/10.1038/nrn3112

Zhang, X., & Samuel, A. G. (2015). The activation of embedded words in spoken word recognition. *Journal of Memory and Language*, *79*, 53–75. https://doi.org/10.1016/j.jml.2014.12.001

Zhang, X., & Holt, L. L. (2018). Simultaneous tracking of coevolving distributional regularities in speech. *Journal of Experimental Psychology: Human Perception and Performance*, *44*(11), 1760–1779. https://doi.org/10.1037/xhp0000569