Journal of Experimental Psychology: Human Perception and Performance

Dimension-Based Statistical Learning of Vowels

Ran Liu and Lori L. Holt

Online First Publication, August 17, 2015. http://dx.doi.org/10.1037/xhp0000092

Journal of Experimental Psychology:
Human Perception and Performance

0096-1523/15/$12.00 http://dx.doi.org/10.1037/xhp0000092

Dimension-Based Statistical Learning of Vowels

Ran Liu and Lori L. Holt
Carnegie Mellon University

Speech perception depends on long-term representations that reflect regularities of the native language. However, listeners rapidly adapt when speech acoustics deviate from these regularities due to talker idiosyncrasies such as foreign accents and dialects. To better understand these dual aspects of speech perception, we probe native English listeners' baseline perceptual weighting of 2 acoustic dimensions (spectral quality and vowel duration) toward vowel categorization and examine how they subsequently adapt to an "artificial accent" that deviates from English norms in the correlation between the 2 dimensions. At baseline, listeners rely relatively more on spectral quality than vowel duration to signal vowel category, but duration nonetheless contributes. Upon encountering an "artificial accent" in which the spectral-duration correlation is perturbed relative to English language norms, listeners rapidly down-weight reliance on duration. Listeners exhibit this type of short-term statistical learning even in the context of nonwords, confirming that lexical information is not necessary to this form of adaptive plasticity in speech perception. Moreover, learning generalizes to both novel lexical contexts and acoustically distinct altered voices. These findings are discussed in the context of a mechanistic proposal for how supervised learning may contribute to this type of adaptive plasticity in speech perception.

Keywords: dimension-based statistical learning, adaptive plasticity, perceptual learning, cue weighting, statistical learning

Human perceptual systems develop stable, long-term representations that reflect the regularities of the environment. Yet, they adapt to short-term deviations in the input. Understanding the balance between maintenance of relatively stable representations and adaptive plasticity is a major challenge for theories of perception. This issue is particularly prominent in speech perception. By adulthood, listeners have formed long-term representations reflecting language-community-specific distributional regularities in the mapping between speech acoustics and linguistic categories (e.g., phonemes). Yet, adult listeners sometimes encounter speech with acoustics that deviate from the norm due to foreign accent, dialects, or speech disorders. For example, a dinnertime conversation may lead a Native Italian speaker to refer to the delicious *chicken* with a vowel more characteristic of English /i/ (a *cheeken*) than /I/ (a *chicken*). Or, a chat about Pittsburgh sports may include reference to the Pittsburgh "*Stillers.*" Due to the local dialect's tense-lax merger, the Pittsburgh Steelers' hometown fans produce the team name with a vowel more similar to English /I/ than /i/ (Johnstone, Andrus, & Danielson, 2006; Labov, Ash, & Boberg, 2005). In each of these cases, the mapping between speech acoustics and the intended message is distorted relative to native English adult listeners' long-term representations.

Systematic distortions like this can reduce intelligibility (Guediche et al., 2014; Mattys et al., 2012). But, repeated exposure to distorted speech in supportive contexts that disambiguate the acoustics (e.g., knowledge that *chicken* is a word, but *cheeken* is not) results in learning that promotes intelligibility and is sustained even in the absence of disambiguating contexts (Bertelson, Vroomen, & de Gelder, 2003; Kraljic et al., 2008; Norris, McQueen, & Cutler, 2003; van Linden & Vroomen, 2007). Moreover, in some cases, learning generalizes to other speech sounds and different talkers (Kraljic & Samuel, 2006, 2007; Reinisch & Holt, 2014). Such learning may support a listener's ability to accommodate speech acoustics of dialects and foreign accents that deviate from long-term language regularities (e.g., Norris et al., 2003). Although lexical and visual disambiguating contexts have been most studied (Samuel & Kraljic, 2009; Vroomen & Baart, 2012), other factors can also drive such adaptive changes. For example, short-term perturbations in the local statistical sampling of speech sounds from perceptual space also lead to adaptive changes in how acoustics are mapped to long-term speech representations (Clayards et al., 2008; Idemaru & Holt, 2011, 2014).

The present studies focus on adaptive plasticity[1] in speech perception related to this latter type of disambiguating information. Multiple, probabilistic acoustic dimensions define speech categories. These dimensions covary and differ in their effectiveness in signaling category membership. Some acoustic dimensions are more reliably diagnostic of category membership and may be

Ran Liu and Lori L. Holt, Department of Psychology, Carnegie Mellon University.

We thank Christi Gomez, Howard Soh, and Rachel Browne for help conducting the experiments. This research was supported by a National Institutes of Health Grant R01DC004674, a National Science Foundation Graduate Research Fellowship, an R. K. Mellon Presidential Fellowship (through the Center for the Neural Basis of Cognition), and a National Institutes of Health training Grant T32GM081760).

Correspondence concerning this article should be addressed to Ran Liu, Department of Psychology, Carnegie Mellon University, Pittsburgh, PA 15201. E-mail: ranliu@cmu.edu

[1] Across different studies and tasks, adaptive shifts in speech perception have been variously termed *perceptual learning*, *adaptation*, *recalibration*, and *re-tuning*, where the terminology is driven primarily by the associated task. Here, we use the phrase *adaptive plasticity* to refer to the broader literature reporting listeners' adjustments to accommodate short-term deviations in speech acoustics. We use the phrase *dimension-based statistical learning* to refer to the specific instantiation of adaptive plasticity in the context of the present manipulations of acoustic dimension correlations (and in the contexts of Idemaru & Holt, 2011, 2014).

perceptually weighted more than other dimensions (e.g., Francis et al., 2008; Holt & Lotto, 2006; Iverson & Kuhl, 1995; Nittrouer, 2004). For example, in English, voice onset time (VOT) and fundamental frequency at vowel onset (F0) covary with voicing categories such that voiced consonants like [b] and [d] are produced with shorter VOTs and lower F0s than are voiceless consonants like [p] and [t] (Kingston & Diehl, 1994; Kohler, 1986). Native English speakers treat VOT as the relatively more reliable, unambiguous signal to voicing (Abramson & Lisker, 1985; Francis et al., 2008; Gordon et al., 1993) in voicing categorization. Nonetheless, F0 informs category membership. Especially when VOT information is ambiguous, listeners are more likely to categorize a sound as voiced if F0 is lower (Abramson & Lisker, 1985; Castleman & Diehl, 1996; Whalen et al., 1993). This pattern of perception highlights listeners' sensitivity to long-term regularities of F0/VOT covariation in English.

Although there are regularities in how acoustic dimensions correlate to define multidimensional phonetic categories, these dimension correlations can vary quite a lot in natural speech input as a function of dialect, accent, and speaker idiosyncrasies. Idemaru and Holt (2011, 2014) tested how listeners adapt in response to such deviations in dimension-correlations. They introduced an artificial "accent" by exposing native English listeners to words (rhymes *beer, pier, deer, tear*) in which the correlation between VOT and F0 was reversed from the relationship typical of English (e.g., higher F0s were paired with voiced stops (*beer, deer*), and lower F0s were paired with voiceless stops [*pier, tier*]). In response to just a few trials of exposure to this artificially accented speech, listeners rapidly down-weighted reliance on F0 such that F0 no longer influenced speech categorization. Idemaru and Holt (2011) interpreted these data to suggest that VOT information served as a reliable teaching signal to orient the relationship of the secondary, F0, dimension to the phonetic categories. These results demonstrate that listeners dynamically track relationships between acoustic dimensions in speech processing. The diagnosticity of an acoustic dimension to phonetic category membership is evaluated relative to changing local regularities between acoustic dimensions rather than as simply a fixed function of its value along the acoustic dimension. Idemaru and Holt referred to this adaptive process as *dimension-based statistical learning*.

So far, evidence of dimension-based statistical learning has been limited to stop consonant categorization (F0 and VOT dimensions; Idemaru & Holt, 2011, 2014; VOT dimension, Clayards et al., 2008). These findings are relevant to some real-life short-term speech signal deviations arising from accents. For example, native English speakers learning Korean use the canonical English relationship of VOT and F0 when producing Korean consonants, even though this relationship is not characteristic of Korean (Kim & Lotto, 2002). This produces non-native accented Korean speech that violates the correlations between dimensions typical of native Korean speech and presents a perceptual challenge for native Korean listeners.

However, VOT is somewhat unique among acoustic speech dimensions because it is closely related to a well-documented auditory perceptual discontinuity that may influence the placement of categories along the dimension (Holt, Lotto, & Diehl, 2004; Kuhl & Miller, 1975; Stevens, 1989). Perhaps because of this, VOT distributions tend to have quite little within-category variation relative to other speech sounds. With somewhat tight constraints on the degree to which VOT may vary in signaling category membership, it is possible that speech stimuli varying in VOT may present a somewhat unusual test case for examinations of adaptive plasticity in speech perception.

Vowels, on the other hand, tend to overlap significantly more in formant frequencies than stop consonants do in VOT (e.g., Hillenbrand et al., 1995; Lisker & Abramson, 1964; Peterson & Barney, 1952). Because of the greater number of potentially ambiguous productions, highly overlapping vowel categories may be especially dependent upon listeners' abilities to make use of short-term statistical regularities to "tune" phonetic categorization. Evidence from computational models of learning has shown that models learn well from distributional statistical information across highly acoustically separable phonetic categories like those distinguished by VOT (McMurray, Aslin, & Toscano, 2009; Vallabha et al., 2007) but decline in performance when categories exhibit greater acoustic overlap the way vowel categories do (Feldman, Griffiths, & Morgan, 2009). This suggests that listeners' long-term, learned vowel representations may be less resolved and more overlapping as well. Furthermore, vowel information constrains word recognition less tightly than does consonant information (Cutler et al., 2000). Thus, there is reason to suspect that the ability to track and use statistical information deviating from listeners' long-term representations of vowel categories could differ significantly from what has been observed for stop consonant categories. This is important because a great deal of foreign accent and dialect information is conveyed by vowels (Labov, 1994); adaptive plasticity in vowel perception is an ecologically significant issue. Adaptation to short-term statistical deviations in the input for categories like vowels, however, has not yet been directly investigated.

Across the present experiments, we had three aims. Our first goal was to examine whether the dynamic, dimension-based statistical learning observed by Idemaru and Holt (2011, 2014) extends to vowels. Our second goal was to investigate whether this type of dimension-based statistical learning occurs in the context of nonlexical items. Dimension-based statistical learning differs from lexically guided adaptive plasticity (e.g., Kraljic & Samuel, 2005; Norris et al., 2003) in that lexical information does not disambiguate the ambiguous speech acoustics; all phonetic possibilities are real words (*deer/tear, beer/pier*). However, dimension-based statistical learning has been observed thus far only in the context of lexical items (Idemaru & Holt, 2011, 2014), so it is not clear whether the learning applies to the mapping of acoustics to specific lexical items or to prelexical representations. Our final goal was to probe the level of representation at which dimension-based statistical learning occurs by investigating the degree of generalization to novel contexts and acoustic items. Idemaru and Holt (2014) showed that dimension-based learning does not generalize from one stop consonant contrast (/b/-/p/) to another (/d/-/t/) and that listeners are able to simultaneously track opposing dimension correlation statistics for the two phonetic contrasts. This pattern of results could mean that dimension-based statistical learning is specific to phonetic categories, the surrounding phonetic context, or the acoustical details of the exemplars experienced in the artificial accent. The present experiments disambiguate the features across which dimension-based statistical learning generalizes.

Toward these aims, we focused on the dimensions of spectral quality and vowel duration in signaling the English /ɛ/-/æ/ vowel contrast. We collected baseline data, by sampling vowels with

values across the full spectral continuum crossed with the full duration continuum, on native English adults' relative perceptual weighting of the two dimensions arising from long-term experience with English (Experiment 1). We then adapted the experimental paradigm of Idemaru and Holt (2011) to investigate whether listeners rapidly adapt dimension-weighting in response to short-term experience with deviations, in the correlation between the spectral quality and duration dimensions, from those typically present in English (Experiment 2). We then replicated Experiment 2 with an artificial accent composed of vowels acoustically identical to those of Experiment 2, but presented in a nonlexical context (Experiment 3). Finally, we assessed whether adaptive dimension reweighting generalizes to acoustically identical vowels that are presented in a different word-frame (Experiment 3) and to acoustically distinct productions of the same vowels by an altered voice (Experiment 4).

Experiment 1

Production data from native English speakers indicates that phonetically distinct but spectrally similar American English vowels (e.g., /i/-/ɪ/, /u/-/ʊ/, /æ/-/ɛ/) are systematically contrastive along both spectral quality and vowel duration dimensions. In English, productions of /ɛ/ tend to be shorter in duration, on average, than productions of /æ/ (Hillenbrand et al., 1995; Hillenbrand et al., 2000). However, spectral quality serves as a relatively better predictor of category membership (Hillenbrand et al., 1995), and accordingly, native English listeners tend to rely more upon this dimension in vowel perception. When both dimensions vary among vowel productions, native English listeners' categorization responses are better predicted by spectral quality than by vowel duration (Hillenbrand et al., 2000; Kondaurova & Francis, 2008). Nevertheless, duration has a significant influence on vowel perception, even as the relatively weaker dimension (Hillenbrand et al., 2000; Kondaurova & Francis, 2008; Ainsworth, 1972).

Although the relative perceptual weighting of spectral quality and duration dimensions to vowel categorization by native English speakers has been investigated for the /i/-/ɪ/ distinction using a stimulus set in which both dimensions are orthogonally manipulated (Kondaurova & Francis, 2008), perceptual dimension weighting for spectral and duration dimensions has not been systematically mapped for /ɛ/-/æ/.

In Experiment 1, native English listeners categorized vowels varying in spectral quality and duration across a two-dimensional stimulus grid. Using the approach of Holt and Lotto (2006), we calculated listeners' relative perceptual weights across the dimensions. These dimension-weights estimate the influence of long-term English experience on perceptual weights and serve as a baseline for interpreting native English listeners' responses to the short-term statistical manipulations (the "artificial accent") introduced in Experiments 2 through 4.

Method

Participants. Forty-seven adults (ages 18 to 30 years) participated in the experiment for either university credit or a small payment. All participants were Carnegie Mellon University students or employees. All were native English speakers and reported normal hearing in both ears.

Stimuli. Forty-nine stimuli were constructed by crossing a seven-step series varying along the spectral quality dimension (from a canonical /ɛ/ to a canonical /æ/) with a seven-step vowel duration series.

The stimuli were created from natural recordings of the words SET and SAT produced by an adult female speaker, with slightly exaggerated vowel duration lengths to serve as stimuli on the "long-duration" end of the vowel duration spectrum. One exemplar of each word was selected based on vowel quality and roughly equivalent vowel durations between the two vowels. These tokens served as endpoints for creating the spectral series at the longest duration value.

The relatively steady-state portions of the vowels /ɛ/ and /æ/ were spliced from their respective words at zero crossings of the waveform and the values of the first four formant trajectories were extracted using Burg's formant extraction algorithm (maximum five formants; maximum formant value = 5500 Hz; 0.025-s time window; preemphasis from 50 Hz) in Praat (Boersma & Weenink, 2009). The values of each of the formant trajectories were interpolated at equal steps between /ɛ/ and /æ/ using R (R Development Core Team, 2008), and these values were entered into Praat to generate the seven-step spectral series. Because this manipulation affected formant frequencies across the entire spectrum by gradually shifting all formants, we refer to this acoustic dimension as spectral quality.

Each of the tokens along this seven-step spectral series was systematically reduced in vowel duration using Praat's PSOLA function to yield identical spectral series varying in vowel duration from 175 ms to 475 ms in 50-ms steps. The most central vowel duration values along this series straddle the boundary between average adult native English-speaking female /ɛ/ and /æ/ production durations measured by Hillenbrand and colleagues (1995).

Each of these /ɛ/-/æ/ vowels was concatenated with an /s/ production preceding the vowel and a /t/ production following the vowel to re-create the words SET and SAT. The /s/ and /t/ segments were isolated productions from the same speaker who generated the original SET and SAT productions from which the vowels were extracted. Each segment's duration was normalized to be equal to its average duration across the contexts of SET and SAT productions. Care was taken to concatenate the consonants and vowels at zero crossings in the waveform. The /s/ and /t/ productions were acoustically identical for each stimulus across the entire vowel stimulus grid.

Procedure. On each trial, participants heard one of the 49 resulting words diotically over headphones as a screen prompted them to choose either SET (by pressing the key *Z*) or SAT (by pressing the key *M*). The relative positioning of the words SET and SAT on the prompt screen (with SET on the left and SAT on the right) matched the relative positioning of the corresponding keys on the keyboard. The visual prompts remained consistently on the screen throughout all trials within a block.

All words were sampled at 44,100 Hz. Participants had an opportunity to make a response as soon as each word file had finished playing; any keystroke made before the full duration of the word had played out was not logged. After participants made a valid response, there was a 1-s pause before the next word was presented. Thus, they progressed through the trials in a self-timed fashion and did not experience any time pressure for making their responses.

There were 10 blocks in the experiment. Each of the 49 stimuli was presented exactly once in each block in randomized order, resulting in 490 total trials. Participants were given the opportunity to take a self-paced break between each block.

Results

The proportion of SAT responses was calculated for each of the 49 stimuli. The results, averaged across all participants, are summarized in a heatmap representation in Figure 1, where the color for each stimulus box in the grid ranges from completely white (no SAT responses) to completely black (all SAT responses). Visual inspection of the heatmap suggests a strong influence of spectral quality on categorization but also an influence of vowel duration, particularly at the most ambiguous spectral values.

To assess the contributions of spectral quality and duration toward listeners' vowel categorization, we conducted a mixed-effects logistic regression analysis with participant as a random effect, spectral quality of the stimulus and duration of the stimulus as fixed effects, and the participant's response (SET or SAT) as the outcome. These analyses reveal that both spectral quality ($p < .001$) and duration ($p < .001$) were significant contributors to participants' categorization responses. The coefficients were 1.062 (standard error = 0.013) for spectral quality and 0.262 (standard error = 0.010) for duration. These results suggest that, although listeners rely more heavily on the spectral dimension, vowel duration significantly influences categorization as well. Furthermore, the influence of vowel duration is most prominent when spectral information is ambiguous, as evidenced by the continuum at Spec = 4 in the heatmap representation (see Figure 1).

Following the approach of Holt and Lotto (2006), perceptual weights for the dimensions were computed for each subject as the correlation between dimension values (either spectral quality or duration) and proportion SAT responses across all stimuli. The absolute values of the correlation coefficients were normalized to sum to one. The relative perceptual dimension weights averaged across participants were 0.822 for spectral quality and 0.178 for duration (standard error for each weight was 0.019). The distribution of individual spectral dimension weights (Figure 1, inset) shows that the vast majority of listeners rely more on spectral quality than duration (the majority of spectral weight values cluster between 0.6 and 1) in /ɛ/-/æ/ vowel categorization. Thus, both average and individual listeners' dimension-weight data converge with visual observations (see Figure 1) that spectral quality was the dominant dimension signaling /ɛ/-/æ/ vowel identity to native listeners.

These baseline perceptual dimension weights are thought to reflect listeners' long-term representations formed from extensive experience with the statistical regularities present in the productions of native language speech (Escudero et al., 2009; Francis et al., 2008; Holt & Lotto, 2006; Iverson et al., 2003; Nittrouer, 2004). Thus, we consider these dimension weights a reflection of native English listeners' long-term representations for /ɛ/ and /æ/.

Experiment 2

With these dimension weights as a baseline, we examine how manipulating the stimulus sampling in short-term input through introduction of an artificial accent alters the relative weighting of these two dimensions in vowel categorization. Similar to the experimental setup of Idemaru and Holt (2011), listeners experience four blocks of trials. In each block, stimuli are drawn from the Experiment 1 stimulus grid defined across spectral quality and duration. The first block samples the grid in a neutral fashion, similar to Experiment 1, to get a baseline measure of the individual listener's dimension weights. In the second and fourth blocks, the majority of trials sample regions of the grid in a manner that mirrors the relationship of spectral quality to duration in American

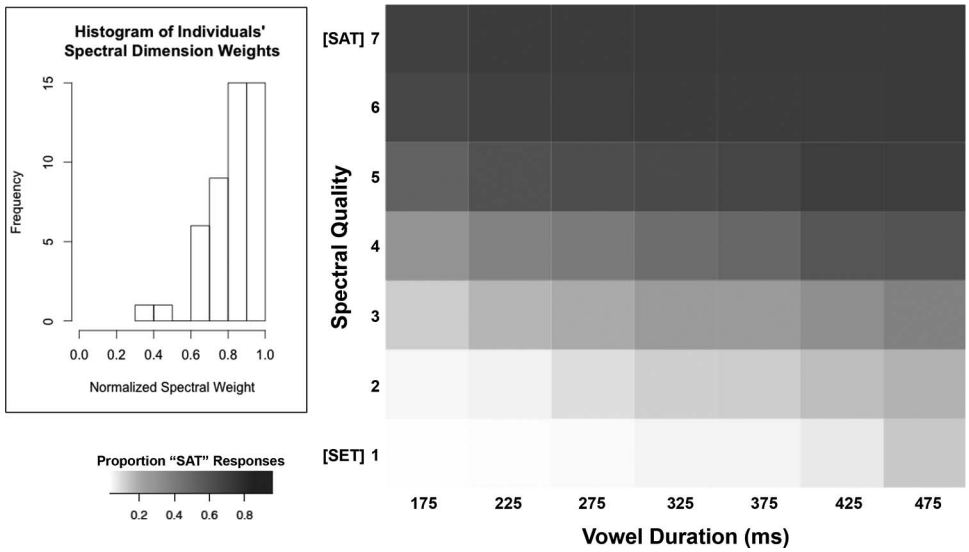

Figure 1. Heat map of proportion SAT (recordings of the word SAT produced by an adult female speaker) responses for each of the 49 stimuli in Experiment 1. Pure white corresponds to no SAT responses; pure red corresponds to all SAT responses. Histogram showing the distribution of all 47 individual listeners' spectral dimension weights, normalized with duration cue weights to sum to 1.

English speech productions. In the third block, the stimuli sample the opposite quadrants of the grid such that the correlation of vowels' spectral quality and duration dimensions is opposite that of canonical English speech. In this way, the experimental design introduces an artificial accent during this third, "Reverse" block. A pair of test trials with long versus short duration but ambiguous spectral quality information appears in each block. These test stimuli provide a means by which to assess listeners' reliance on the duration dimension in vowel categorization as the short-term speech regularities change across blocks.

Although this is an artificial laboratory-created "accent," it is consistent with the kind of acoustic dimension-based deviations that listeners may experience in natural accented speech. For example, although the 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). These relationships are reflected in perception among listeners from these language communities; native Scottish–English listeners perceptually weight spectral information more than listeners from Southern-English backgrounds. Thus, the speech of a Southern English talker is characterized by a correlation between spectral quality and duration dimensions that runs counter to a Scottish English listener's long-term experience.

In the present experiment, the Reverse block reverses the correlation of vowels' spectral quality and duration dimensions relative to American English speech (similar to the experimental setup used in Idemaru & Holt, 2011 for stop consonant categorization). In a controlled laboratory setting, this mimics the shift in statistical regularities away from long-term expectations as, for example, when a Scottish listener encounters a Southern English talker. In Experiment 2, we examine the impact of this dimension-correlation manipulation on listeners' reliance on duration to signal vowel category.

Method

Participants. Twenty adults (ages 18 to 30 years) participated in the experiment for either university credit or a small payment. All participants were either university students or employees. All were native English speakers and reported normal hearing in both ears.

Stimuli. The sounds in the experiment were selectively sampled from the grid of 49 stimuli described in Experiment 1. The stimulus sampling depended on the specific block of the experiment, described in greater detail below and summarized in Figure 2.

Procedure. Across each of four blocks, native-English listeners heard the words SET and SAT and responded with a key press to indicate which word they heard. Within each block, there were exposure trials and test trials, intermixed with order of presentation randomized. On exposure trials, vowels' spectral quality robustly signaled vowel identity across all blocks. The correlation between vowels' spectral quality and duration, however, varied across blocks to be neutral, consistent with, or inconsistent with the correlation typical of native English speech. There were also test trials covertly embedded in all blocks. The two test trial stimuli had identical spectral energy, held constant at the stimulus step whose spectral quality was most ambiguous with respect to signaling vowel identity based on results from Experiment 1, but differed in vowel duration. These stimuli are depicted in Figure 2 as the diamonds and squares, which represent the short-duration (225 ms) and long-duration (425 ms) test stimuli, respectively. Details of the stimulus sampling for the exposure trials in each of the four blocks are as follows:

Pretest block. Participants were exposed to the 25 most central stimuli of the 49-stimulus grid (see Figure 2a). The purpose of the Pretest block was to provide a neutral sampling of the stimulus grid such that each value across the spectral quality dimension was paired with each value across the duration dimension. Similar to Experiment 1, this provides a neutral dimension-correlation environment to collect baseline data on the two test stimuli of interest (while minimizing the duration of the block with 25, instead of 49, unique stimuli). Within each block, participants heard 10 repetitions of the full set of exposure stimuli, for a total of 250 trials. Note that this stimulus set included the two test stimuli.

Canonical block. Participants heard a selective sampling of 18 stimuli that were consistent with the familiar, canonical correlation of vowel duration and spectral quality found in American English speech (see Figure 2b). The sampling included nine stimuli with /ɛ/-like spectral energy and short vowel durations (Figure 2b, lower left corner) and another nine stimuli with /æ/-like spectral information and long vowel durations (Figure 2b, upper right

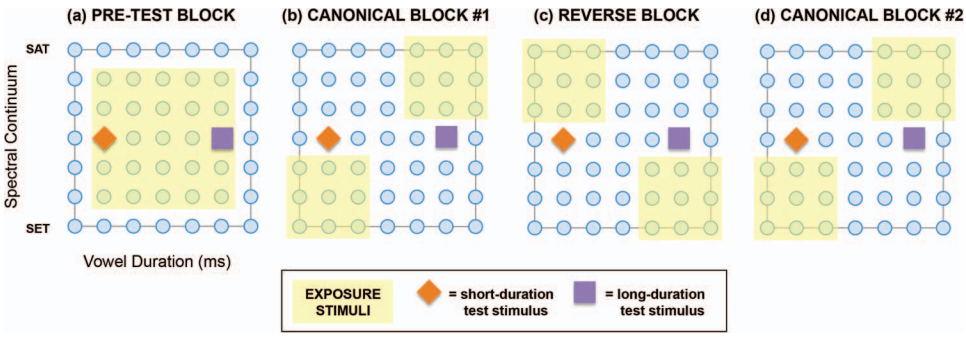

Figure 2. Summary of the selective sampling of exposure stimuli (boxes) and test stimuli (diamonds and squares) for each of the four experimenter-defined blocks in Experiments 2 through 4. Note that, in the Pretest block, the exposure stimulus set included the two test stimuli so they were not included in any trials separate from the regular exposure trials. See the online article for the color version of this figure.

corner). The two test stimuli (Figure 2b, diamond and square) were also included in the block. Participants heard 10 repetitions of set of 18 exposure and two test stimuli for a total of 200 trials.

Reverse block. Participants heard a selective sampling of 18 stimuli that exhibited a correlation between vowel duration and spectral quality opposite from the canonical relationship found in English (see Figure 2c). The sampling included nine stimuli with /ɛ/-like spectral energy and long vowel durations (Figure 2c, lower right corner) and another nine stimuli with /æ/-like spectral information and short vowel durations (Figure 2c, upper left corner). The two test stimuli (Figure 2c, diamond and square) were also included in the block. Participants heard 10 repetitions of set of 18 exposure and two test stimuli for a total of 200 trials. Note that spectral quality, the dimension relied upon more strongly by the native English listeners of Experiment 1, robustly signaled vowel identity but its relationship with the duration dimension ran counter to long-term English experience in this artificial accent.

The experimenter-defined blocks were ordered such that every participant received them in the following order: Pretest, Canonical, Reverse, Canonical. The duration of each block ranged from approximately 10 to 15 min, depending on how quickly the participant was responding to each trial. We hypothesized that, if the rapid dimension weight changes observed by Idemaru and Holt (2011, 2014) hold true for these vowel categories, listeners will down-weight reliance on the duration dimension in the Reverse block. Since test trials are characterized by perceptually ambiguous spectral quality information but distinct durations, vowel categorization of the two test trials serves as an index of reliance upon duration in vowel categorization. Following the approach of Idemaru and Holt (2011, 2014), we included a repetition of the Canonical block to differentiate potential mechanisms. If listeners continue to track the spectral/duration correlation throughout the Reverse block and beyond, we predict that the effect of duration on spectrally ambiguous test stimuli will rebound in the final block when the regularity shifts back to the canonical correlation. If, however, listeners remap the relationship in a more permanent manner or cease to be sensitive to duration, more generally, as a result of the artificial accent in the Reverse block then the effects should persist into the second Canonical block. The participant instructions and procedure were identical to those of Experiment 1.

Results

Dimension weights. To ensure that spectral quality was indeed the dominant dimension signaling /ɛ/-/æ/ vowel identity for the listeners in this experiment, we computed dimension weights from pretest block data using the same method as in Experiment 1. The relative perceptual dimension weights averaged across participants were 0.705 for spectral quality and 0.295 for duration (standard error for each weight was 0.054). These weights confirm that spectral quality was the dominant dimension among these specific participants.

Categorization of exposure stimuli. Results from Experiment 1 and the above analyses demonstrate that spectral quality is a strong signal of vowel category. Therefore, we predicted highly accurate vowel categorization of exposure trials, which were relatively unambiguous in spectral quality across all blocks of the experiment. Indeed, this was the case. The proportion of exposure stimuli that were categorized as expected based on the spectral dimension was 0.944 (standard error = 0.028) in the first Canonical block, 0.848 (standard error = 0.039) in the Reverse block, and 0.933 (standard error = 0.032) in the second Canonical block. The highly accurate vowel categorization confirms that spectral quality robustly signaled vowel identity.

Categorization of test stimuli. Figure 3a shows how the local statistics within an experiment block affected vowel categorization of the test stimuli.[2] To analyze the effect of different exposure blocks on categorization of the two different test stimuli, a 4 (Block: Pretest, Canonical 1, Reverse, Canonical 2) × 2 (Duration of test stimulus: long, short) repeated-measures analysis of variance (ANOVA) was run on arcsine-transformed proportion SAT responses. Results revealed a significant main effect of duration, $F(1, 19) = 48.44$, $p < .0001$, and a significant interaction between block and duration, $F(3, 57) = 30.75$, $p < .0001$. There was no significant main effect of block. The significant Block × Duration interaction indicates that the influence of duration on vowel categorization differed across blocks. To explore the basis of this interaction, planned *t* tests comparing arcsine-transformed proportion SAT responses to long- versus short-duration test stimuli were run with a Bonferroni-adjusted alpha of 0.0125 (to correct for multiple comparisons). These revealed that duration exerted a significant effect on categorization for the Pretest, $t(19) = 5.579$, $p < .0001$; Canonical 1, $t(19) = 7.975$, $p < .0001$; and Canonical 2, $t(19) = 8.514$, $p < .0001$, blocks. Duration did not significantly affect categorization for the Reverse block, $t(19) = 0.111$, $p = .913$. The mean difference scores in proportion SAT responses between the long- versus short-duration test stimuli (Figure 3b) were 0.47 ($SE = 0.073$) for Pretest, 0.56 ($SE = 0.055$) for Canonical 1, 0.01 ($SE = 0.058$) for Reverse, and 0.58 ($SE = 0.061$) for Canonical 2. These difference scores suggest a marked decrease in reliance on duration to signal vowel category in the Reverse block when the spectral/duration dimension correlation experienced across exposure stimuli was reversed.

These analyses methods were chosen to parallel those used in previous studies on dimension-based statistical learning (Idemaru & Holt, 2011, 2014). Acknowledging issues with using ANOVA to analyze proportion data computed from categorical outcomes (Jaeger, 2008), we also analyzed the categorization responses using mixed logit models (generalized linear mixed model for binomially distributed outcomes). We constructed a mixed logit model that was driven by our hypothesis that listeners exhibit a significant down-weighting of reliance on the duration dimension in the Reverse block. To this end, we modeled the outcome—participants' categorization responses (with SET coded as 0 and SAT coded as 1) as a function of participant, block, test stimulus duration (long, short), and the interaction between block and test stimulus duration. Participant was modeled as a random effect, whereas the other factors were modeled as fixed effects. Results of this model revealed a significant interaction between block and test

[2] Error bars are not plotted in graphs presenting responses to different stimuli that are both within-subject and within-condition. The consistency of across-condition differences, not within-condition variability, contributes to statistical significance and so error bars on these graphs can be visually misleading. Error bars are plotted in graphs representing difference scores computed from within-subject conditions because, in these cases, error bars indicating the variability of difference scores are appropriate and meaningful.

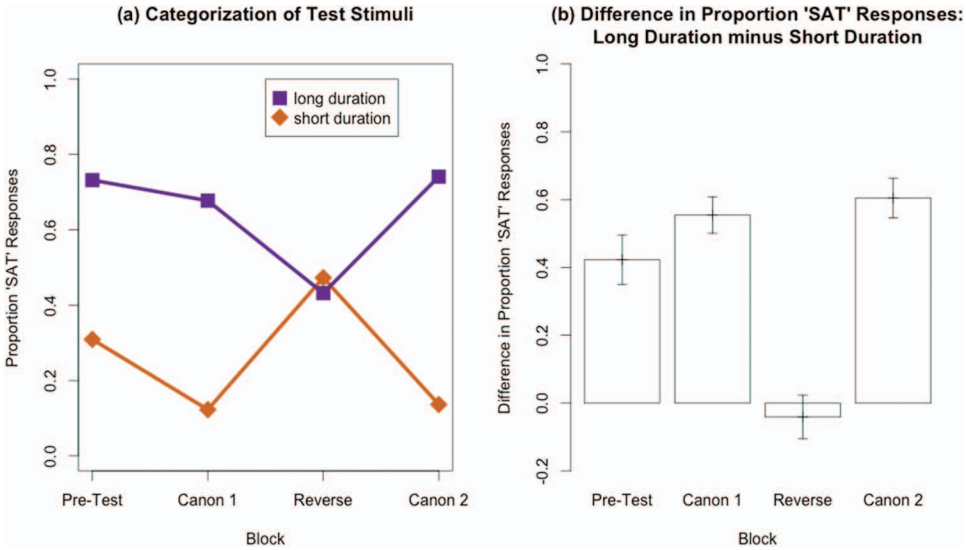

Figure 3. Experiment 2 results. (a) Proportion SAT responses, organized by block and test stimulus (long vs. short duration). (b) Duration-driven differences in proportion SAT responses. Calculated by subtracting, for each individual participant, the proportion SAT responses for the short-duration test stimulus from that for the long-duration test stimulus. Error bars represent standard error of the mean. See the online article for the color version of this figure.

stimulus duration ($\beta = 0.093$, $SE = 0.019$, $p < .0001$). To explore the basis of this interaction, we modeled the effect of test stimulus duration separately for each of the four blocks (modeling participant as a random effect and test stimulus duration as a fixed effect). There was a significant effect of test stimulus duration in the first block (Pretest: $\beta = 0.409$, $SE = 0.065$, $p < .0001$), in the second block (Canonical 1: $\beta = 0.761$, $SE = 0.081$, $p < .0001$), and in the fourth block (Canonical 2: $\beta = 0.797$, $SE = 0.078$, $p < .0001$). There was no significant effect of test stimulus duration in the third block (Reverse: $\beta = -0.048$, $SE = 0.030$, $p = .109$).

The results from all of our analyses converge to show that vowel duration significantly affected categorization in the blocks in which the correlation between spectral quality and duration was either neutral or consistent with regularities of English (Pretest, Canonical 1, and Canonical 2 blocks). In contrast, listeners rapidly down-weighted reliance on the duration dimension when the local, short-term speech environment was characterized by an "artificial accent" that reversed the correlation between spectral quality and duration relative to that typical of English. In the Reverse block, the very same test-trial vowels that were differentiated as a function of duration in the Canonical blocks were no longer as different vowels by listeners.

Discussion

These results demonstrate that listeners are sensitive to the correlation between spectral quality and duration dimensions contributing to the /ɛ/-/æ/ vowel distinction and adjust perceptual dimension weighting in response to short-term deviations in the correlation. Experiencing a temporary reversal of the canonical correlation between spectral quality and duration dimensions in the Reverse block led to a down-weighting of the contribution of duration in differentiating spectrally ambiguous vowels. Results from Experiment 1 as well as categorization data from the current experiment confirm that the spectral dimension very robustly signals vowel category on the exposure trials (the majority of experiment trials). Thus, we hypothesize that listeners may dynamically adjust the reliability of the duration cue to vowel identity when its correlation to the dominant spectral cue is no longer consistent with long-term representations. We return to this possibility in the General Discussion.

A notable aspect of these data is that when the speech input returned to the typical dimension correlation in the second Canonical block, duration again contributed to vowel categorization. Therefore, the perceptual down-weighting of duration cannot be attributed to a complete inattention to the duration dimension. For this pattern to emerge, duration must have been processed, at some level, throughout the Reverse block, although it ceased to reliably signal vowel identity. This pattern of findings is indicative of a highly sensitive and dynamically responsive perceptual system adapting to the statistical regularities of incoming speech. Results from this experiment very closely mirror the pattern of results observed in Idemaru and Holt (2011), suggesting that the type of short-term adjustment observed for the relationship of VOT and F0 dimensions in signaling stop consonants generalizes to dimensions distinguishing a vowel contrast.

Experiment 3

In the dimension-based statistical learning demonstrated in Experiment 2, both vowel choices (SET, SAT) form real words. Because of this, presumably lexical information is not of use in disambiguating the vowel or in driving the adaptive plasticity we observe. However, dimension-based statistical learning of this sort has only been observed in the context of lexical items (Experiment 2; Idemaru & Holt, 2011, 2014), so it is not clear whether learning

plays out in the mapping of acoustics to specific lexical items or to prelexical units. Having used only word recognition to examine this effect, one possibility is that short-term deviations in the statistical regularity of dimensions signaling a particular word specifically affect the mapping from input dimension to lexical item. If this were the case, we would expect the down-weighting of duration to depend on exposure to real words and to evoke word-specific effects that fail to generalize to other words possessing the same vowel. In the present experiment, we probe whether dimension-based statistical learning occurs if listeners are exclusively exposed to statistical deviations of acoustic dimensions in the context of nonwords. To this end, we used the same dimension-correlation manipulations as in Experiment 2 with the same vowel stimuli embedded in a nonlexical context (SETCH/SATCH). Further, we introduced infrequent generalization stimuli with vowels acoustically identical to the SETCH/SATCH test stimuli but with a different—and lexical—consonant-frame context (SET/SAT). This allowed us to assess whether adaptive changes in perceptual dimension-weighting in response to statistical deviations in a set of vowels in one context (S_TCH) generalizes to acoustically identical vowels present in a different context (S_T).

Method

Participants. Nineteen adults (ages 18 to 22 years) participated in the experiment for either university credit or a small payment. All participants were university students, native English speakers and reported normal hearing in both ears.

Stimuli. The vowels in the experiment were selectively sampled from the grid of 49 stimuli described in Experiment 1. The vowel sampling across blocks was identical to the sampling in Experiment 2 (see Figure 2), and the vowels were acoustically identical to those used in Experiments 1 and 2. The only difference was that the vowels were in a 'S_TCH' frame such that stimuli would either be perceived as the nonword SETCH or the nonword SATCH.

The stimuli were created by removing the /t/ segments from each of the 49 SET/SAT productions created for Experiment 1 (and used in Experiment 2) and replacing them with /tch/ segments. The /tch/ segment was produced in isolation by the same speaker who recorded the original SET/SAT stimuli. Care was taken to cut the /t/ segments and insert the /tch/ segments at zero crossings in the waveform. The /tch/ segments were acoustically identical across the entire stimulus grid.

Procedure. The trial structure, within-block sampling of vowels, structure of blocks across the experiment, and participant instructions were identical to those for Experiment 2. The main difference in the present experiment is that listeners heard the exposure trial vowels in the context of the nonwords SETCH and SATCH rather than SET and SAT. The vowel segments of all exposure stimuli, however, were identical to the exposure stimuli in Experiment 2. Embedded within every block were two regular test trials with vowels that matched those of the Experiment 2 test trials (identical spectral quality at the most ambiguous value, but differing in vowel duration) but in the nonword frame S_TCH to match the exposure stimuli. Also embedded within each block were two generalization test trials which were the same vowels as the regular test trials but presented in the word frame S_T. These generalization test trials were included to assess whether listeners' dimension-based statistical learning in the context of the nonlexical items SETCH and SATCH would generalize to lexical items like SET and SAT. On each trial, listeners were given four orthographic response options to choose from: SETCH, SATCH, SET, and SAT. We included all four options on every trial to keep all trials as similar as possible and to avoid drawing attention to the distinction between exposure and test trials.

Results

Dimension weights. To ensure that spectral quality was indeed the dominant dimension signaling /ɛ/-/æ/ vowel identity for the listeners in this experiment, we computed dimension weights from Pretest block data using the same method as in Experiment 1. The relative perceptual dimension weights averaged across participants were 0.733 for spectral quality and 0.267 for duration (standard error for each weight was 0.021). These weights confirm that spectral quality is the dominant dimension among these specific participants.

Categorization of exposure stimuli. Vowel categorization of the SETCH/SATCH exposure stimuli in the Canonical and Reverse blocks was highly accurate. The proportion of exposure stimuli that were categorized as expected based on the spectral dimension was 0.942 ($SE = 0.017$) in the first Canonical block, 0.932 ($SE = 0.022$) in the Reverse block, and 0.96 ($SE = 0.011$) in the second Canonical block. The high categorization performance across all blocks indicates that spectral quality was a robust signal of vowel identity.

Categorization of test stimuli. Figure 4 plots the results. To analyze the effect of exposure blocks on categorization of the four different test stimuli, a 4 (Block: Pretest, Canonical 1, Reverse, Canonical 2) × 2 (Duration of test stimulus: long, short) × 2 (Test stimulus type: regular S_TCH, generalization S_T) repeated-measures ANOVA was run on arcsine-transformed proportion /æ/ responses (either SATCH or SAT depending on the test stimulus type). Results revealed a significant main effect of duration, $F(1, 18) = 36.6$, $p < .0001$]; a significant main effect of block, $F(1, 18) = 11.17$, $p < .0001$; and a significant interaction between block and duration, $F(3, 57) = 38.9$, $p < .0001$. There was no significant main effect of test stimulus type (whether they were regular S_TCH or generalization S_T), and there were no significant interactions between it and any other variables. The significant Block × Duration interaction indicates that the influence of duration on vowel categorization differed across the four blocks.

To explore the basis of this interaction among the regular (S_TCH) test stimuli, planned t tests comparing arcsine-transformed proportion SATCH responses to long- versus short-duration test stimuli were run with a Bonferroni-adjusted alpha of 0.0125 (to correct for multiple comparisons). Duration exerted a significant effect on categorization in the Pretest, $t(18) = 5.904$, $p = .0001$; Canonical 1, $t(18) = 6.168$, $p < .0001$; and Canonical 2, $t(18) = 7.361$, $p < .0001$, blocks but not in the Reverse block, $t(18) = 1.052$, $p = .307$. This pattern of results for the regular test stimuli is summarized in Figure 4a. The mean difference scores in proportion SATCH responses between the long- versus short-duration test stimuli were 0.28 ($SE = 0.042$) for Pretest, 0.51 ($SE = 0.072$) for Canonical 1, −0.09 ($SE = 0.075$) for Reverse, and 0.55 ($SE = 0.063$) for Canonical 2, exhibiting a steep decrease in reliance on duration to differentiate the vowels the Reverse

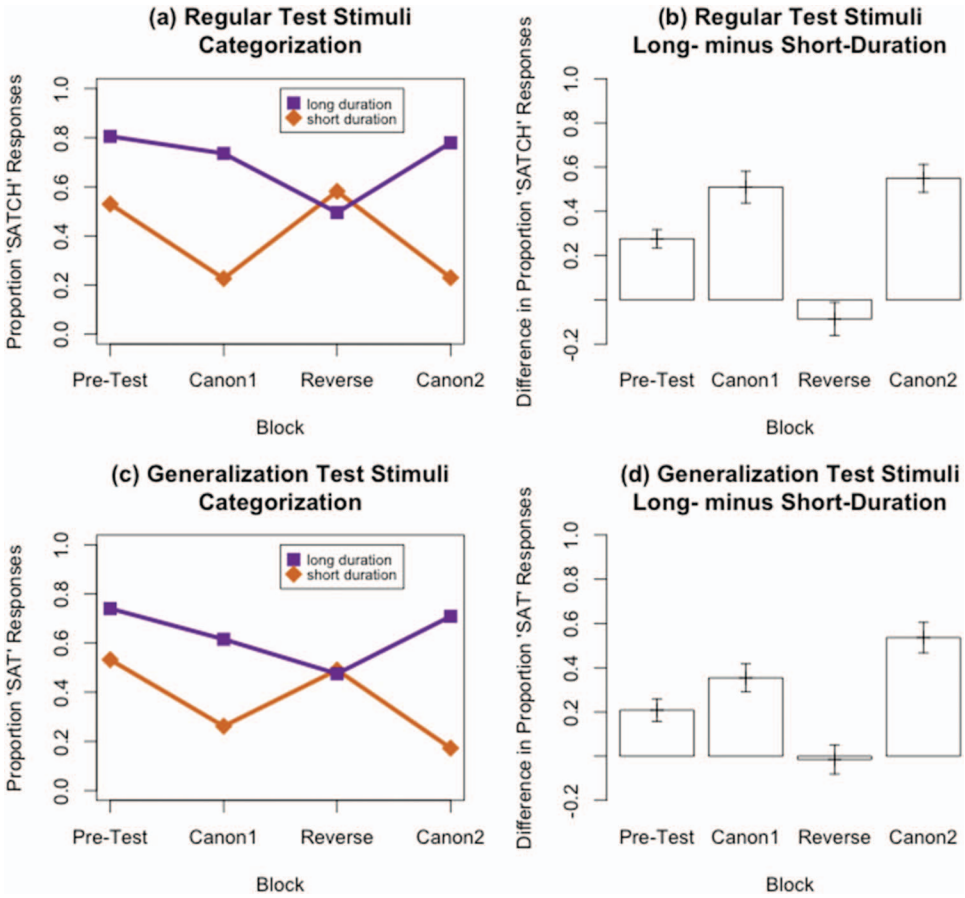

Figure 4. Experiment 3 results. (a) Proportion SATCH responses, organized by block and test stimulus (long vs. short duration) for regular test stimuli (SETCH/SATCH). (b) Duration-driven differences in proportion SATCH responses. (c) Proportion SAT responses for generalization test stimuli (SET/SAT). (d) Duration-driven differences in proportion SAT responses. See the online article for the color version of this figure.

block when the exposure stimuli were sampled such that the spectral/duration dimension correlation was reversed. These difference scores are plotted in Figure 4b.

To explore the basis of this interaction among the generalization (S_T) test stimuli, planned *t* tests comparing arcsine-transformed proportion SAT responses to long- versus short-duration test stimuli were run with a Bonferroni-adjusted alpha of 0.0125 (to correct for multiple comparisons). Duration exerted a significant effect on categorization for the Pretest, $t(18) = 4.03$, $p = .0008$; Canonical 1, $t(18) = 5.286$, $p < .0001$; and Canonical 2, $t(18) = 6.881$, $p < .0001$, blocks but not for the Reverse block, $t(18) = 0.86$, $p = .399$. The pattern of results for the Generalization test stimuli is summarized in Figure 4c. The mean difference scores in proportion SAT responses between the long- versus short-duration test stimuli were 0.21 ($SE = 0.051$) for Pretest, 0.35 ($SE = 0.063$) for Canonical 1, −0.02 ($SE = 0.065$) for Reverse, and 0.54 ($SE = 0.069$) for Canonical 2, demonstrating a decrease in reliance on duration to differentiate the SET/SAT test stimuli in the Reverse block when the spectral/duration dimension correlation was reversed among the SETCH/SATCH exposure stimuli. These difference scores are plotted in Figure 4d.

We also analyzed the categorization responses using mixed logit models. We modeled the participant's categorization response (with SETCH/SET coded as 0 and SATCH/SAT coded as 1) as a function of participant, block number, test stimulus duration (long, short), test stimulus type (regular, generalization), and interactions between the latter three factors. Participant was modeled as a random effect, whereas the other factors were modeled as fixed effects. Results of this model revealed a significant interaction between block and test stimulus duration ($\beta = 0.085$, $SE = 0.026$, $p = .0009$) and no other statistically significant interactions. To further investigate the significant Block × Duration interaction, we modeled the effect of test stimulus duration separately for each of the four blocks (modeling participant as a random effect and test stimulus duration as a fixed effect). There was a significant effect of test stimulus duration in the first block (Pretest: $\beta = 0.277$, $SE = 0.038$, $p < .0001$), in the second block (Canonical 1: $\beta = 0.510$, $SE = 0.049$, $p < .0001$), and in the fourth block (Canonical 2: $\beta = 0.698$, $SE = 0.054$, $p < .0001$). There was no significant effect of test stimulus duration in the third block (Reverse: $\beta = -0.024$, $SE = 0.031$, $p = .426$).

To summarize, the analyses converge to show that vowel duration significantly affected categorization in the blocks for which the correlation between spectral quality and duration was neutral (Pretest) or consistent with that of English (Canonical blocks). In the Reverse block in which exposure trials sampled the vowel

acoustic space to create an artificial accent with the opposite dimension correlation of English, duration no longer affected vowel categorization. Thus, we observed that dimension-based statistical learning occurs even when the artificial accent is experienced only in the context of nonwords. It does not appear that the adaptive plasticity is occurring at the level of the mapping to specific lexical representations. We acknowledge the possibility that, with exposure to the nonwords SETCH and SATCH across the four blocks, it is possible that these nonwords may have begun to take on some word-like qualities. This would, however, be quite a weak view of what it means for an item to be "lexical" as these nonwords are not linked to any semantic representations, referents, or other perceptual input.

The present pattern of dimension-based statistical learning was observed not only for the regular test stimuli (S_TCH) context but also the generalization test stimuli (S_T), with no significant difference in the pattern of results based on test stimulus type. Thus, adaptation to the artificially "accented" speech fully generalized to a word-frame context that was never heard spoken in the artificial accent.

Experiment 4

Experiment 3 demonstrated that dimension-based statistical learning generalizes across small differences in surrounding phonetic contexts. Listeners experiencing this learning on the /ɛ/-/æ/ vowel contrast in the context of nonwords SETCH and SATCH readily generalized to vowels in the context of the words SET and SAT. Idemaru and Holt (2014) showed that listeners do not generalize adaptive perceptual dimension weighting from one stop consonant contrast (/b/-/p/) to another (/d/-/t/). Instead, the adaptive plasticity seems to either be specific to the phonetic contrast itself or the acoustic details of the exemplars experienced in the artificial accent across exposure trials.

Here, we test the possibility that dimension-based statistical learning is specific to the acoustic details of the exemplars across which listeners experience an artificial accent. Specifically, we investigate whether listeners' dimension-based learning from artificially accented /ɛ/-/æ/ vowels generalizes to acoustically distinct exemplars of those same vowels.

Method

Participants. Nineteen adults (ages 18 to 22 years) participated in the experiment for either university credit or a small payment. All participants were university students. All were native English speakers and reported normal hearing in both ears.

Stimuli. The vowels in the experiment were selectively sampled from the grid of 49 stimuli described in Experiment 1. The sampling of exposure stimuli across blocks was identical to the sampling in Experiment 2 (see Figure 2) and, like in Experiment 2, all vowels were situated in a S_T frame.

A separate grid of 49 acoustically distinct generalization vowel stimuli was generated by applying the automated "change gender" function in Praat to each of the original stimuli. The spectral content of each generalization test stimulus was substantially different from that of the corresponding regular test stimulus (see Figure 5 for a formant comparison of the spectrally ambiguous stimuli for regular and generalization stimulus sets).

To ensure that native English listeners' categorization of /ɛ/-/æ/ within the generalization vowel grid elicited perceptual dimension weights similar to those of the training (and regular test) stimulus grid, we collected a small group (n = 10) of native English listeners' categorization responses for the full 49-stimulus generalization grid. Indeed, the relative perceptual weights (calculated according to the methods described in Experiment 1) for spectral quality (0.728) and duration (0.272) dimensions were similar to those reported in Experiment 1 for the training (and regular test) stimuli (0.822 for spectral and 0.178 for duration). Furthermore, the effect of vowel duration on perception of the spectrally ambiguous tokens was robust for the generalization stimuli and revealed a categorization curve highly similar to that observed for the original training (and regular) test stimuli.

Procedure. The trial structure, within-block sampling of vowels, structure of blocks across the experiment, and participant instructions were identical to those for Experiment 2. The difference from Experiment 2 was that there were two acoustically-distinct generalization test trial stimuli in addition to the two regular test trial stimuli (the same as used in Experiment 2). The two generalization test trial stimuli were also spectrally ambiguous tokens, with durations equal to those of the regular test trials. However, only the spectral acoustic details of the regular test trials matched those of the exposure trials across blocks, whereas the spectral acoustic details of the generalization test trials were substantially different.

Results

Dimension weights. To ensure that spectral quality was the dominant dimension signaling /ɛ/-/æ/ vowel identity for the listeners in this experiment, we computed dimension weights based on participants' categorization of the exposure stimuli (same acoustics as regular test stimuli) in the Pretest block. Dimension weights were calculated using the same method as in Experiment 1. The relative perceptual dimension weights averaged across participants were 0.767 for spectral quality and 0.233 for duration (standard error for each weight was 0.022). These weights confirm that spectral quality was the dominant dimension among these specific participants.

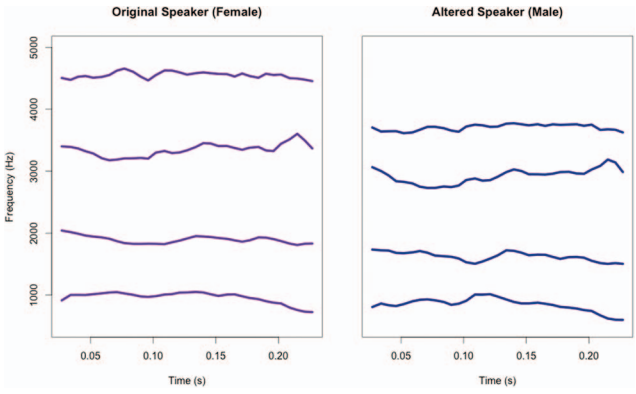

Figure 5. Extracted formant tracks for the spectrally ambiguous vowels in the original stimulus grid (from which exposure trials and regular test trials are drawn) [left/purple] and the generalization stimulus grid [right/blue]. See the online article for the color version of this figure.

Categorization of exposure stimuli. Vowel categorization of the exposure stimuli in the Canonical and Reverse blocks was highly accurate. The proportion of exposure stimuli that were categorized as expected based on the spectral dimension was 0.968 (standard error = 0.007) in the first Canonical block, 0.954 (standard error = 0.015) in the Reverse block, and 0.978 (standard error = 0.006) in the second Canonical block. Accurate categorization across blocks indicates that spectral quality was a robust signal of vowel identity.

Categorization of test stimuli. Results are shown in Figure 6. To analyze the effect of different exposure blocks on categorization of the four different test stimuli, a 4 (Block: Pretest, Canonical 1, Reverse, Canonical 2) × 2 (Duration of test stimulus: long, short) × 2 (Test stimulus type: regular, generalization) repeated-measures ANOVA was run on arcsine-transformed proportion SAT responses. Results revealed a significant main effect of duration, $F(1, 18) = 157.1$, $p < .0001$; and test stimulus type, $F(1, 18) = 8.126$, $p = .011$; a significant interaction between block and duration, $F(3, 54) = 35.82$, $p < .0001$; and a significant interaction between block, duration, and test stimulus type, $F(3, 54) = 9.202$, $p < .0001$. The significant block by duration interaction indicates that the influence of duration on vowel categorization was dependent on the block, and the significant block by duration by test stimulus type interaction indicates that this influence differed for regular test stimuli as compared to generalization (acoustically distinct) test stimuli.

We first explored the basis of the Block × Duration interaction among the regular test stimuli. Planned t tests comparing arcsine-transformed proportion SAT responses to long- versus short-duration test stimuli were run with a Bonferroni-adjusted alpha of 0.0125 (to correct for multiple comparisons). Results showed that duration exerted a significant effect on categorization in the Pretest, $t(18) = 7.088$, $p < .0001$; Canonical 1, $t(18) = 12.202$, $p < .0001$; and Canonical 2, $t(18) = 13.527$, $p < .0001$; blocks but did not affect categorization for the Reverse block, $t(18) = 0.054$, $p = .957$. This pattern of results for the regular test stimuli is summarized in Figure 6a. The mean difference scores in proportion SAT responses between the long- versus short-duration regular test stimuli were 0.36 ($SE = 0.052$) for Pretest, 0.68 ($SE = 0.05$) for Canonical 1; −0.01 ($SE = 0.068$) for Reverse; and 0.72 ($SE =$

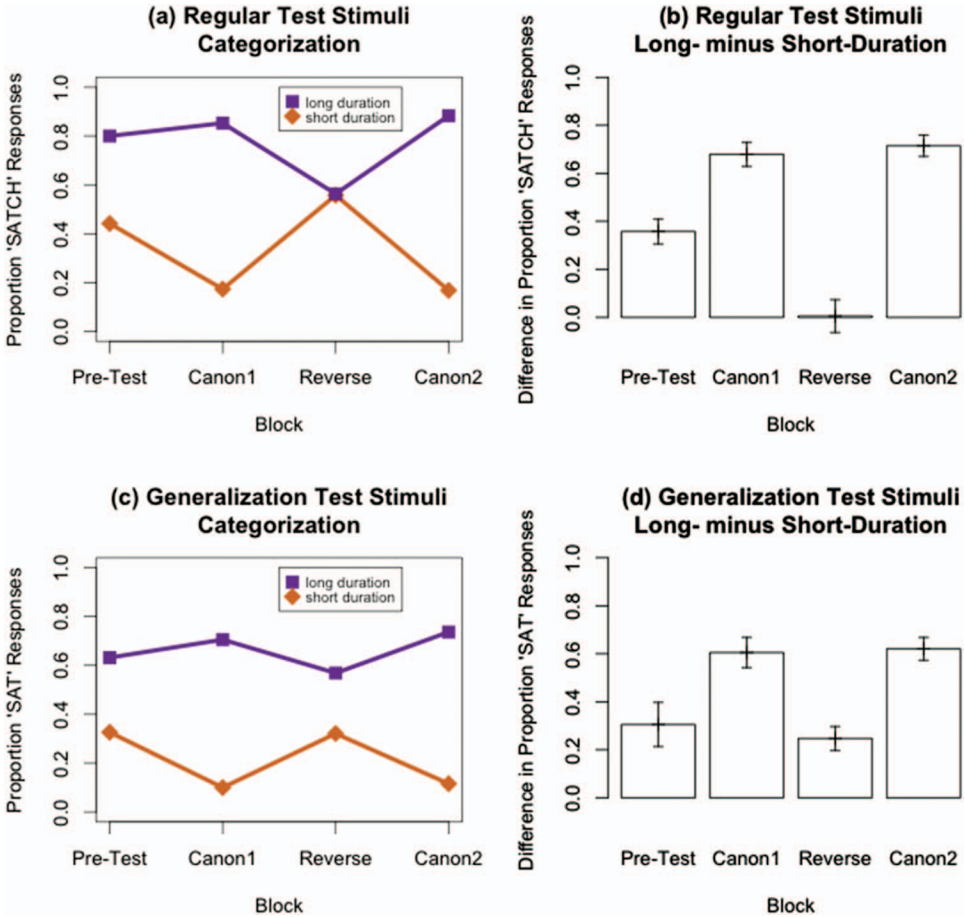

Figure 6. Experiment 4 results. (a) Proportion SAT responses, organized by block and test stimulus (long vs. short duration) for regular test stimuli (same speaker as in the exposure stimuli). (b) Duration-driven differences in proportion SAT responses for regular test stimuli. (c) Proportion SAT responses for generalization test stimuli (different speaker from exposure stimuli). (d) Duration-driven differences in proportion SAT responses for generalization test stimuli. See the online article for the color version of this figure.

0.044) for Canonical 2, charting a steep decrease in reliance on duration to differentiate the test stimuli in the Reverse block when the spectral/duration dimension correlation among exposure stimuli was reversed. These difference scores are plotted in Figure 6b.

Furthermore, *t* tests revealed that duration exerted a significant effect (with significance tested at a Bonferroni-adjusted alpha of 0.0125) on categorization for the generalization (acoustically distinct) test stimuli for all four blocks: Pretest, $t(18) = 3.346$, $p = .0036$; Canonical 1 [$t(18) = 7.696$, $p < .0001$; Canonical 2 [$t(18) = 11.303$, $p < .0001$]; and Reverse, $t(18) = 4.749$, $p = .0002$. To assess whether listeners significantly down-weighted duration in the Reverse block relative to in the Canonical blocks, a 4 (Block: Pretest, Canonical 1, Reverse, Canonical 2) × 2 (Duration of test stimulus: long, short) ANOVA was conducted solely on the arcsine-transformed proportion SAT responses to generalization stimuli. Results showed a significant main effect of duration, $F(1, 18) = 77.78$, $p < .0001$, and a significant interaction between block and duration, $F(3, 54) = 9.333$, $p < .0001$. The significant interaction between block and duration, considered along with the *t* test results, suggests that even for the generalization test stimuli listeners relied on duration less so in the Reverse block than in the other blocks, even though reliance on duration remained significant in the Reverse block. This pattern of results for the generalization test stimuli is summarized in Figure 6c. Furthermore, the mean difference scores in proportion SAT responses between the long- versus short-duration generalization test stimuli were 0.31 ($SE = -0.092$) for Pretest, 0.61 ($SE = 0.062$) for Canonical 1, 0.25 ($SE = 0.049$) for Reverse, and 0.62 ($SE = 0.047$) for Canonical 2, indicative of a significant decrease in the effect of duration on vowel categorization in the Reverse block. These difference scores are plotted in Figure 6d.

We also analyzed the categorization responses using mixed logit models. We modeled the participant's categorization response (with SET coded as 0 and SAT coded as 1) as a function of participant, block number, test stimulus duration (long, short), test stimulus type (regular, generalization), and interactions between the latter three factors. We modeled participant as a random effect and the other factors as fixed effects. Results of this model revealed a significant interaction between block and test stimulus duration ($\beta = 0.155$, $SE = 0.014$, $p < .0001$), a significant interaction between block, test stimulus duration, and test stimulus type ($\beta = -0.036$, $SE = 0.007$, $p < .0001$), and no other significant interactions. To further investigate the significant three-way (block, test stimulus duration, test stimulus type) interaction, we modeled the effects of test stimulus duration, test stimulus type, and their interaction within each of the four separate blocks (also modeling participant as a random effect). There was a significant effect of test stimulus duration in Block 1 (Pretest: $\beta = 0.338$, $SE = 0.080$, $p < .0001$), Block 2 (Canonical 1: $\beta = 0.790$, $SE = 0.074$, $p < .0001$), and Block 4 (Canonical 2: $\beta = 0.878$, $SE = 0.080$, $p < .0001$). In these three blocks, there was no significant interaction between test stimulus duration and test stimulus type. In Block 3 (Reverse), there was a significant simple effect of test stimulus duration ($\beta = 0.262$, $SE = 0.054$, $p < .0001$) and a significant interaction between test stimulus duration and test stimulus type (regular vs. generalization; $\beta = -0.256$, $SE = 0.075$, $p = .0006$). These results align with previous *t* test results to suggest that, in the Reverse block, the effect of duration on categorization of acoustically distinct generalization test stimuli is reduced relative to that in the Canonical blocks but not as reduced as that for the regular test stimuli (whose acoustics are much more similar to the exposure stimuli).

Thus, listeners' categorization of the generalization (acoustically distinct) test stimuli mirrored that of the regular test stimuli in all blocks except for the Reverse block, where the duration down-weighting was significant relative to the Canonical blocks, but less extreme. Duration continued to significantly impact vowel categorization of generalization stimuli even in the Reverse block, although its influence was significantly down-weighted relative to the Canonical blocks. The regular test stimuli (those that acoustically match the exposure stimuli), however, led to an extreme down-weighting of duration, replicating our observations in previous experiments. This provides evidence of generalization of dimension-based statistical learning to vowels that belong to the same category as, but whose acoustic details are quite distinct from, the vowels experienced in the "artificial accent". However, the degree of generalization appears to be modulated by the acoustic similarity across vowels.

General Discussion

One of the greatest challenges in studying perceptual systems is to understand how they dynamically adapt to short-term deviations in the input while simultaneously maintaining stable long-term representations. Speech perception is a good test-bed for this issue because adult listeners have established long-term representations that reflect native language regularities, yet they adapt to short-term deviations from these norms (e.g., in foreign accented speech, dialect, etc.). There has been mounting evidence that listeners rely on various information sources to adaptively adjust speech categorization (e.g., Clayards et al., 2008; Eisner & McQueen, 2006; Holt, 2005; Idemaru & Holt, 2011; Kraljic & Samuel, 2006, 2007; Norris et al., 2003). The present experiments demonstrate that listeners adjust speech categorization to accommodate artificially "accented" speech in which dimension correlations deviate from those in typical English productions of vowels /ɛ/ and /æ/ and capitalize on this finding to better understand the mechanisms involved in adaptive plasticity in speech perception.

Experiment 1 established the relative perceptual weights that native English listeners assign to two acoustic dimensions, spectral quality and duration, in categorizing /ɛ/ and /æ/. Consistent with expectations from prior research, listeners treat spectral quality as the most diagnostic cue to vowel categorization for this contrast. When spectral quality is ambiguous, though, listeners do rely significantly on duration. There was considerable consistency in this pattern across participants. Experiment 2 demonstrated that listeners track the relationship between these two dimensions in online speech perception as short-term distributional regularities shift to create an artificial "accent." Specifically, listeners rapidly down-weight reliance on duration for vowel categorization when the correlation between spectral quality and duration is reversed relative to the long-term English norm. Test stimuli with highly distinct stimulus durations, but ambiguous spectral quality, signaled opposing vowel categories in the Pretest and Canonical blocks. However, these exact same vowels were indistinguishable in the Reverse block. The very same vowel acoustics signaled vowel categorization differently as a function of the statistical regularities characterizing local speech experience in the block.

These findings serve as a conceptual replication of the results reported in Idemaru and Holt (2011, 2014) and extend the pattern of results to vowel categories.

An additional aim of the present work was to understand the mechanistic basis of these effects. Prior work investigating dimension-based statistical learning has involved only real lexical items in training. This leaves open the possibility that the observed dimension down-weighting may be word-specific, affecting only the mapping to a particular lexical representation. Another unresolved issue was whether adaptation occurs in a highly specific manner that only impacts experienced tokens. We find evidence contrary to each of these possibilities. In Experiment 3, listeners exposed only to nonword tokens in the artificial accent exhibited the same pattern of duration down-weighting as observed in Experiment 2. Moreover, this learning generalized to word tokens not heard spoken in the artificial accent. Although these data rule out the possibility that the dimension-based statistical learning is word-specific, the word and nonword tokens in Experiment 3 shared identical vowel acoustics. In Experiment 4, we tested the possibility that the effect is specific to the vowel acoustics experienced in the artificial accent. The results indicate that dimension-based statistical learning generalizes to same-category (/ɛ/-/æ/) vowels with different acoustics. However, the down-weighting of duration in the Reverse block was attenuated for the acoustically differing novel exemplars compared to the acoustically identical trained exemplars. This suggests that the degree of acoustic similarity between trained and generalization exemplars influences the degree to which dimension-based statistical learning generalizes. This is of interest in that the influence of acoustic similarity on generalization has also been observed in the context of another form of adaptive plasticity in speech perception, lexically guided phonetic tuning (Kraljic & Samuel, 2005; Reinisch & Holt, 2014).

Mechanistic Proposal

We propose that the present results may be accounted for by positing a multilevel representational network with assumptions similar to interactive activation models like TRACE (Mirman et al., 2006; McClelland & Elman, 1986) in which the initial connection weights among representations are related to perceptual weights, learned through long-term regularities in the perceptual environment. The baseline dimension weights we collected across equally balanced, orthogonally varying dimensions in Experiment 1 can approximate the relative strength of initial connection weights in such a network. Experiment 1 revealed a stronger baseline dimension weight for spectral quality relative to vowel duration. This suggests strong initial connection weights that more effectively signal vowel category via the spectral quality dimension and weaker connection weights to the same categories from the duration dimension.

To accommodate the pattern of adaptive plasticity evident in Experiments 2 to 4, the model's connection weights would need to be modifiable. Prior modeling efforts (Hebb-TRACE, Mirman et al., 2006) have incorporated Hebbian learning that modifies connection weights to account for lexically mediated adaptive plasticity in speech categorization. Although this approach could simulate the present pattern of dimensional reweighting, Hebbian learning is likely to be too sluggish to account for the rapid learning we observe in the present studies and in Idemaru and Holt (2011, 2014; see Guediche et al., 2014 for discussion). We have speculated that supervised learning mechanisms may be more aligned with the time course of behavioral learning we observe (Guediche et al., 2014; Idemaru & Holt, 2011) because tasks in which an internal model of a target representation can be used to detect and correct for differences from actual experience can tap into error-based supervised learning (see Guediche et al., 2014). In this context, the relatively unambiguous spectral quality of exposure trials may serve as a "teaching signal" that is sufficient to robustly activate vowel categories based on strong connections weights established by long-term experience. We hypothesize that this allows the system to derive a prediction of the expected duration via the long-term mapping of duration to vowel category. These predictions may be compared with the actual sensory input, with any 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. We speculate that the specific adjustments could occur via one of two different mechanisms.

One possibility is that an error signal could destabilize the connection weights of duration to vowel category uniformly across the duration dimension. In other words, the contribution of duration as a dimension may be dampened, or "ignored," at least at the category decision level. This mechanism is supported by the fact that listeners did not exhibit a reversal of mapping between duration values and vowel categories in the Reverse block (e.g., with longer durations signaling /ɛ/ and shorter durations signaling /æ/). Such a reversal would be expected if listeners were simply learning to mirror the statistics of the short-term environment. The idea that listeners stopped treating duration, more generally, as a cue to vowel categorization, is consistent with the pattern of results observed here. This dimension down-weighting, however, is not occurring at the level of dimension encoding. We observe a rapid change in duration-weighting when the dimension correlation statistics change back to canonical following the Reverse block. This indicates that listeners were still encoding duration information in the Reverse block, even if it was not contributing to category decisions.

An alternative mechanism that may support the behavior we observe is that listeners' connection weights may adjust to better match the input–output relationship, for example strengthening the connection weights between short vowel durations and the /æ/ category and between long vowel durations and the /ɛ/ category. These strengthened connection weights may lead to different balances in category-level activation, which would affect how competitive dynamics play out at the category level. As an example, consider a vowel production with ambiguous spectral quality between /ɛ/ and /æ/ and a long vowel duration. In this case, based on long-term representations, the ambiguous spectral quality would not be effective at activating vowel categories in a manner that would differentiate /ɛ/ and /æ/. But, a long duration would tip the balance by more robustly activating the /æ/ vowel category. If experience with artificially accented speech strengthens the connection between long vowel durations and the /ɛ/ category (the reverse correlation), then the test trials ambiguous in spectral quality with long durations may begin to activate the /æ/ and /ɛ/ categories more uniformly than prior to exposure to the accent. This would effectively balance their activation of /ɛ/ and /æ/ categories, leading to a less clear-cut "winner" in the battle of

inhibitory dynamics at the category level. Activating both category alternatives in this way for both short- and long-duration test stimuli may result in a "neutral" effect of duration, as observed behaviorally in the Reverse blocks of Experiments 2 to 4. This could be a way for learning systems to avoid allowing the adjustment to short-term statistical regularities that deviate from the norm to override the perceptual system's long-term representations.

The present data do not differentiate between the dimension-destabilization and competition-based accounts. Each of these possibilities is consistent with computational-level (Marr & Poggio, 1979) modeling of adaptation effects in speech (Kleinschmidt & Jaeger, 2015) to describe what the system does and why it does these things. Understanding the mechanistic bases of adaptive processes in speech will require additional next-generation experiments to differentiate the effects at the algorithmic or representational level (Marr & Poggio, 1979). These experiments will need to target questions of specifically how the system does what it does and which representations and processes are involved. For example, extending listeners' exposure to the statistics of the Reverse block (perhaps over the course of days or weeks) ought to lead to a stronger reversal of mapping between duration and vowel category if the competition-based account is correct, whereas there should be no effect of extending exposure if listeners are simply destabilizing the duration dimension relationship to vowel categories. Prior experiments reveal that exposing listeners to a reversal in dimension correlations across five days is not sufficient to reverse the mapping in voicing categories (Idemaru & Holt, 2011), but the time course of exposure may have been insufficient. Vowel categorization may provide a fruitful test-bed for examining these possibilities since fine-grain within-category manipulations of duration and spectral quality are feasible. The present data thus provide a critical foundation for further mechanistic examinations of adaptive plasticity in speech perception.

The results of Experiments 3 and 4 further specify the details of how such a representational network might respond to speech that differs from the speech experienced in the accent. In Experiment 3 the down-weighting of vowel duration was evoked by nonwords and generalized to words never experienced in the artificial accent. This suggests that the representations against which input is compared are at the vowel category level, and not the lexical level. Results from Idemaru and Holt (2014) showed that dimension-based statistical learning did not generalize from one phonetic contrast (/b/-/p/) to another (/d/-/t/), despite the fact that the phoneme pairs were from the same class of sounds (stop consonants) and are typically contrasted together as voiceless (/p, t/) versus voiced stops (/b, d/). The specificity of learning observed by Idemaru and Holt could have been due either to the specific phonetic categories across which the artificial accent was experienced, or to the acoustic details of the exemplars conveying the accent. In Experiment 4, listeners' reliance on vowel duration decreased significantly in response to artificially accented speech in the Reverse block. This was true even for vowel category exemplars with acoustic details distinct from those experienced in the artificial accent. Nonetheless, the degree of down-weighting was less for these acoustically distinct vowels than that observed for acoustically identical vowels. In the context of the mechanistic proposal sketched above, this may suggest that the change in connection weights between the duration dimension and vowel categories relies on similarity to the spectral quality of the vowel exemplars experienced in the artificial accent. Dimension-based learning might not equally affect the relationship between duration and all exemplars of the /ɛ/ and /æ/ vowels.

Similar findings related to the impact of acoustic similarity on generalization have been reported for other forms of adaptive plasticity in speech perception. In the context of lexically guided phonetic retuning, short-term phonetic retuning generalizes to phonemes from previously unheard speakers, but only when exposure and test speakers' utterances are drawn from a similar acoustic or perceptual space (Reinisch & Holt, 2014; Kraljic & Samuel, 2005). In this regard, the substantial within-category variability in vowel acoustics presents an excellent testing ground. In future studies, it will be informative to carefully manipulate the acoustic and perceptual similarity of exemplars experienced in the accent and generalization stimuli to gain further insight regarding the mechanisms of learning involved. Similarly, an interesting line of future research would be to test whether listeners can simultaneously track distinct cue-correlation statistics for the same vowels spoken by a novel, highly acoustically distinct talker. If such mechanisms are to support listeners' rapid adaptation to accents and speech idiosyncrasies, which are often speaker-specific, we expect listeners should be able to track simultaneous statistics. Idemaru and Holt (2014) present evidence that listeners can simultaneously track distinct cue-correlation statistics in different stop-consonant phoneme contrasts (e.g., /b/-/p/ and /d/-/t/), but demonstrating a similar ability for different talkers would be a valuable extension of this work.

Relation to Long-Term Representations

The mechanistic account described above suggests that the effectiveness of the acoustic dimension serving as a "teaching signal" relates to its initial connection weights to the vowel categories, established by long-term experience. This presents the possibility that there may be individual differences in dimension-based statistical learning arising from differences in long-term representations. Preliminary correlation analyses across participants in Experiments 2–4 support this possibility. There was a significant negative correlation between individual listeners' perceptual weight for the dominant (spectral quality) dimension and the effect of duration on vowel categorization observed in response to the artificial accent in the Reverse block (measured as the difference between long- and short-duration test stimuli categorization; $R = -0.4$, $p = .001$). Even after removing the two individuals who weight vowel duration relatively more strongly than spectral quality at baseline (and thus may have been outliers), a significant correlation remained ($R = -0.38$, $p = .003$). Individuals with greater reliance on spectral quality in the baseline test tended to down-weight duration more, resulting in smaller differences in categorization of test stimuli varying only in vowel duration.

The vast majority of native English listeners, and the majority of participants in the present studies, more strongly weight spectral quality in /ɛ/-/æ/ categorization. This makes it difficult to sample across the entire range of potential variability in perceptual cue-weighting to fully test this possibility in the present data. A recent study (Schertz et al., 2013) may be telling in this regard. Owing to language patterns in Korean, VOT and F0 each serve as reliable cues to stop consonant category membership and listeners can be classified as to whether they perceptually weight F0 or VOT most robustly in baseline tests of consonant categorization. These two groups of lis-

teners show quite different patterns of dimension-based statistical learning even with exposure to identical stimulus sets across blocks. In each group, the more dimension weighted most strongly at baseline serves as the more effective teaching signal; the relatively "weaker" perceptual dimension is down-weighted upon exposure to an artificial accent that reverses the typical dimension correlation (Schertz et al., 2013). Investigating a similar relationship between initial dimension weights and dimension-based statistical learning could be accomplished for vowel spectral quality and duration by recruiting participants from native language backgrounds for which duration is a better predictor of vowel category than spectral quality, like Finnish (Ladefoged & Maddieson, 1996). On the basis of our mechanistic account of dimension-based statistical learning, listeners who weight duration more toward vowel categorization would be expected to rely on duration as the "teaching signal." Thus, if presented with the same Reverse block, we would expect that they would learn to down-weight spectral quality as a signal of vowel categorization.

Demonstrations of adaptive plasticity in speech categorization, whereby disambiguating information tunes the way that speech perception interprets ambiguous speech acoustics, are rapidly growing in literatures that span multiple paradigms (see Guediche et al., 2014 for review). As such, these findings contribute to a broader literature examining the balance between the stability of long-term representations and the dynamic adjustment of perception to accommodate short-term deviations from long-term norms. Consonant with the model we sketch above, research on lexically- (Norris et al., 2003) and visually guided (e.g., Vroomen & Baart, 2012; Vroomen et al., 2007) adaptive plasticity in speech perception has proposed mechanisms that are consistent with a supervisory error-signal driven account. However, whereas lexical and visual information serve as supervisory signals guiding adaptation or learning in these paradigms, the present studies establish that neither is necessary to observe adaptive plasticity in speech perception. The supervisory signal in the present effects instead originates from one of the input dimensions (spectral quality) that reliably signals vowel category membership, even in the artificial accent. An exciting possibility for future research will be to examine the extent to which common learning mechanisms drawing off of different sources of information (visual, lexical, acoustic) and interacting with different long-term representations evoke adaptive plasticity in speech perception.

References

Abramson, A. S., & Lisker, L. (1985). Relative power of cues: F0 shift versus voice timing. In V. Fromkin (Ed.), *Phonetic linguistics: Essays in honor of Peter Ladefoged* (pp. 25–32). Orlando, FL: Academic Press.

Ainsworth, W. A. (1972). Duration as a cue in the recognition of synthetic vowels. *The Journal of the Acoustical Society of America, 51* (2B), 648–651. http://dx.doi.org/10.1121/1.1912889

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

Boersma, P., & Weenink, D. (2009). Praat: Doing phonetics by computer (Version 5.1. 20). [Computer program]. Retrieved from http://www.praat.org/

Castleman, W. A., & Diehl, R. L. (1996). Effects of fundamental frequency on medial and final voice judgments. *Journal of Phonetics, 24,* 383–398. http://dx.doi.org/10.1006/jpho.1996.0021

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

Cutler, A., Sebastián-Gallés, N., Soler-Vilageliu, O., & van Ooijen, B. (2000). Constraints of vowels and consonants on lexical selection: Cross-linguistic comparisons. *Memory & Cognition, 28,* 746–755. http://dx.doi.org/10.3758/BF03198409

Eisner, F., & McQueen, J. M. (2006). Perceptual learning in speech: Stability over time. *The Journal of the Acoustical Society of America, 119,* 1950–1953. http://dx.doi.org/10.1121/1.2178721

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. J. Do, L. Domínguez, & A. Johansen (Eds.), *Proceedings of the 25th annual Boston University conference on language development*. Somerville, MA: Cascadilla Press.

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

Feldman, N. H., Griffiths, T. L., & Morgan, J. L. (2009). The influence of categories on perception: Explaining the perceptual magnet effect as optimal statistical inference. *Psychological Review, 116,* 752–782. http://dx.doi.org/10.1037/a0017196

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

Gordon, P. C., Eberhardt, J. L., & Rueckl, J. G. (1993). Attentional modulation of the phonetic significance of acoustic cues. *Cognitive Psychology, 25,* 1–42. http://dx.doi.org/10.1006/cogp.1993.1001

Guediche, S., Holt, L. L., Laurent, P., Lim, S. J., & Fiez, J. A. (2014). Evidence for cerebellar contributions to adaptive plasticity in speech perception. *Cerebral Cortex, 25,* 1867–1877.

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

Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *The Journal of the Acoustical Society of America, 97,* 3099–3111. http://dx.doi.org/10.1121/1.411872

Holt, L. L. (2005). Temporally nonadjacent nonlinguistic sounds affect speech categorization. *Psychological Science, 16,* 305–312. http://dx.doi.org/10.1111/j.0956-7976.2005.01532.x

Holt, L. L., & Lotto, A. J. (2006). Cue weighting in auditory categorization: Implications for first and second language acquisition. *The Journal of the Acoustical Society of America, 119* (Pt. 1), 3059–3071. http://dx.doi.org/10.1121/1.2188377

Holt, L. L., Lotto, A. J., & Diehl, R. L. (2004). Auditory discontinuities interact with categorization: Implications for speech perception. *The Journal of the Acoustical Society of America, 116,* 1763–1773. http://dx.doi.org/10.1121/1.1778838

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

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

Iverson, P., & Kuhl, P. K. (1995). Mapping the perceptual magnet effect for speech using signal detection theory and multidimensional scaling. *The Journal of the Acoustical Society of America, 97,* 553–562. http://dx.doi.org/10.1121/1.412280

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

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards Logit mixed models. *Journal of Memory and Language, 59,* 434–446.

Johnstone, B., Andrus, J., & Danielson, A. E. (2006). Mobility, indexicality, and the enregisterment of "Pittsburghese." *Journal of English Linguistics, 34,* 77–104. http://dx.doi.org/10.1177/0075424206290692

Kim, M. R. C., & Lotto, A. J. (2002). An investigation of acoustic characteristics of Korean stops produced by non-heritage learners. *The Korean Language in America, 7,* 177–188.

Kingston, J., & Diehl, R. L. (1994). Phonetic knowledge. *Language, 70,* 419–454. http://dx.doi.org/10.1353/lan.1994.0023

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,* 148–203. http://dx.doi.org/10.1037/a0038695

Kohler, K. J. (1986). Parameters of speech rate perception in German words and sentences: Duration, F0 movement, and F0 level. *Language and Speech, 29* (Pt. 2), 115–139.

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,* 3959–3971. http://dx.doi.org/10.1121/1.2999341

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

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

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

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

Kuhl, P. K., & Miller, J. D. (1975). Speech perception by the chinchilla: Voiced-voiceless distinction in alveolar plosive consonants. *Science, 190,* 69–72. http://dx.doi.org/10.1126/science.1166301

Labov, W. (1994). *Principles of language change: Internal factors*. Oxford, England: Blackwell.

Labov, W., Ash, S., & Boberg, C. (2005). *The atlas of North American English: Phonetics, phonology and sound change*. Berlin, Germany: Walter de Gruyter. http://dx.doi.org/10.1515/9783110206838

Ladefoged, P., & Maddieson, I. (1996). *Sounds of the world's languages*. Oxford, England: Blackwell.

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

Marr, D., & Poggio, T. (1979). A computational theory of human stereo vision. *Proceedings of the Royal Society of London B: Biological Sciences, 204*(1156), 301–328.

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://dx.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.

McMurray, B., Aslin, R. N., & Toscano, J. C. (2009). Statistical learning of phonetic categories: Insights from a computational approach. *Developmental Science, 12,* 369–378. http://dx.doi.org/10.1111/j.1467-7687.2009.00822.x

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

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

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

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,* 175–184. http://dx.doi.org/10.1121/1.1906875

R Development Core Team. (2008). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org

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

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

Schertz, J. L., Lotto, A. J., Warner, N., & Cho, T. (2013). Acoustic cue weighting across modalities in a non-native sound contrast. *The Journal of the Acoustical Society of America, 134,* 4030–4030. http://dx.doi.org/10.1121/1.4830715

Stevens, K. N. (1989). On the quantal nature of speech. *Journal of Phonetics, 17,* 3–45.

Vallabha, G. K., McClelland, J. L., Pons, F., Werker, J. F., & Amano, S. (2007). Unsupervised learning of vowel categories from infant-directed speech. *Proceedings of the National Academy of Sciences, USA of the United States of America, 104,* 13273–13278. http://dx.doi.org/10.1073/pnas.0705369104

van Linden, S., & Vroomen, J. (2007). Recalibration of phonetic categories by lipread speech versus lexical information. *Journal of Experimental Psychology: Human Perception and Performance, 33,* 1483–1494. http://dx.doi.org/10.1037/0096-1523.33.6.1483

Vroomen, J., & Baart, M. (2012). Phonetic recalibration in audiovisual speech. In M. M. Murray & M. T. Wallace (Eds.), *The neural bases of multisensory processes. Boca Raton,* FL: CRC Press.

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

Whalen, D. H., Abramson, A. S., Lisker, L., & Mody, M. (1993). FO gives voicing information even with unambiguous voice onset times. *The Journal of the Acoustical Society of America, 93* (Pt. 1), 2152–2159. http://dx.doi.org/10.1121/1.406678

Received December 6, 2014
Revision received May 10, 2015
Accepted May 16, 2015 ■