AMERICAN PSYCHOLOGICAL ASSOCIATION

Journal of Experimental Psychology:
Human Perception and Performance

ISSN: 0096-1523

2022, Vol. 48, No. 9, 913–925
https://doi.org/10.1037/xhp0001037

Phonetic Category Activation Predicts the Direction and Magnitude of Perceptual Adaptation to Accented Speech

Yunan Charles Wu and Lori L. Holt
Department of Psychology, Carnegie Mellon University

Unfamiliar accents can systematically shift speech acoustics away from community norms and reduce comprehension. Yet, limited exposure improves comprehension. This perceptual adaptation indicates that the mapping from acoustics to speech representations is dynamic, rather than fixed. But, what drives adjustments is debated. Supervised learning accounts posit that activation of an internal speech representation via disambiguating information generates predictions about patterns of speech input typically associated with the representation. When actual input mismatches predictions, the mapping is adjusted. We tested two hypotheses of this account across consonants and vowels as listeners categorized speech conveying an English-like acoustic regularity or an artificial accent. Across conditions, signal manipulations impacted which of two acoustic dimensions best conveyed category identity, and predicted which dimension would exhibit the effects of perceptual adaptation. Moreover, the strength of phonetic category activation, as estimated by categorization responses reliant on the dominant acoustic dimension, predicted the magnitude of adaptation observed across listeners. The results align with predictions of supervised learning accounts, suggesting that perceptual adaptation arises from speech category activation, corresponding predictions about the patterns of acoustic input that align with the category, and adjustments in subsequent speech perception when input mismatches these expectations.

Public Significance Statement
When we encounter talker with a foreign accent or a stuffy nose, speech is shifted relative to the norm and comprehension can suffer. However, listeners can rapidly adapt to shifts like these by relying on clues from a talker's voice or face, or word knowledge. This study demonstrates that the extent to which listeners adapt to accents and other speech distortions is predicted by the degree to which supporting clues are effective in activating internal representations of speech sound categories, providing insight into the mechanistic basis for the flexibility of speech communication.

Keywords: speech perception, dimension-based statistical learning, adaptive plasticity, perceptual recalibration, perceptual adaptation

Cognitive and perceptual systems face a tension. On the one hand, there is a demand to maintain fairly stable representations that respect long-term environmental regularities. On the other, there is a need to adapt to distinctive short-term regularities as they arise. This pressure is present in speech perception because we are tuned to long-term speech regularities that characterize our dominant language community. Yet, we must routinely adapt when speech departs from these regularities. Whether encountering a voice distorted by a respiratory infection or an English speaker with an unfamiliar accent or dialect, systematic shifts in speech input can negatively impact speech comprehension (Bradlow & Bent, 2008; Clarke & Garrett, 2004). Yet, a bit of experience with a systematic acoustic shift can be sufficient for comprehension to improve, and these improvements sometimes generalize to other contexts with similar shifts to speech acoustics (Bradlow & Bent, 2008; Clarke & Garrett, 2004; Davis et al., 2005; Samuel & Kraljic, 2009; Srinivasan & Zahorik, 2013).

Local Pittsburghers, for example, pronounce their home football team, the Steelers, with /I/ (as in *hill*) rather than /i/ (as in *heel*); thereby, departing from the mapping of acoustics to vowel categories typical of standard American English. But, in the context of a conversation about the upcoming football game, this acoustic ambiguity is resolved by word knowledge: *Steelers* is a team name, *Stillers* is not. Research demonstrates that repeated exposure to ambiguous acoustic speech input in the context of disambiguating

This article was published Online First July 18, 2022.
Yunan Charles Wu https://orcid.org/0000-0002-2243-0760
Lori L. Holt https://orcid.org/0000-0002-8732-4977
The research was supported by a Grant from the National Science Foundation (BCS1941357) to Lori L. Holt and Yunan Charles Wu and by a National Institutes of Health grant to Lori L. Holt (R21DC019217). The authors thank Christi Gomez for her assistance. Study materials and code can be found at Open Science Framework (OSF.io; https://osf.io/m68zd/). This study was not preregistered.
Correspondence concerning this article should be addressed to Lori L. Holt, Department of Psychology, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, United States. Email: loriholt@cmu.edu

information (like the lexical context of *Steelers*) can result in adjustments in speech perception that persist even after the disambiguating information is no longer available (Bertelson et al., 2003; Kraljic & Samuel, 2005, 2006; Norris et al., 2003; Vroomen et al., 2007). In the Steelers example, experiencing the vowel /I/ in word contexts that instead suggests an /i/ can lead listeners to later perceive isolated vowels with ambiguous acoustics between /I/ and /i/ more often as /i/ than /I/.

There is not clear consensus on what drives these adjustments. One account of this perceptual adaptation posits a role for a form of supervised learning driven by disambiguating input (Bertelson et al., 2003; Guediche et al., 2014, 2016; Idemaru & Holt, 2011; Norris et al., 2003). By this view, information in the speech signal that disambiguates systematic shifts in speech acoustics—for instance, lexical in the *Steelers* example (Davis et al., 2005; Norris et al., 2003; Schwab et al., 1985), visual information from articulating faces (Bertelson et al., 2003; Vroomen et al., 2007), orthographic feedback (Guediche et al., 2016; Schwab et al., 1985), or unambiguous acoustic speech cues (Idemaru & Holt, 2011; Liu & Holt, 2015)—resolves the mapping of ambiguous speech acoustics to a particular internal representation, like a phonetic category. In doing so, it may generate expectations of category-typical speech input such that if the ambiguous acoustic input is a poor match to predictions, a mismatch or error signal will drive perceptual adaptation that results in subsequent shifts in speech categorization that remain apparent even when the disambiguating information is no longer present (see Guediche et al., 2014, for a review of this perspective).

By this supervised learning account, activation of an internal representation—a phonetic category in the present studies—is essential in driving adjustments in how acoustic speech input contributes to speech categorization. Two specific hypotheses emerge from this perspective: (a) the effectiveness of phonetic-category-level activation by disambiguating input should relate to the *magnitude* of perceptual adaptation observed and (b) the *directionality* of perceptual adaptation should be malleable according to the information source that disambiguates category membership in a particular listening context; when an acoustic dimension best signals category identity (driving activation), it will result in perceptual adaptation across secondary dimensions. Thus, manipulations that impact which acoustic dimension best signals category identity correspondingly will affect which acoustic dimension is impacted by perceptual adaptation.

Here, we test these two predictions in the context of perceptual adaptation driven by statistical regularities in acoustic speech input, so-called *dimension-based statistical learning* (Idemaru & Holt, 2011, 2014, 2020; Lehet & Holt, 2017, 2020; Liu & Holt, 2015; Schertz et al., 2016; Zhang & Holt, 2018; Zhang et al., 2021). This paradigm elicits perceptual adaptation through systematic shifts in speech acoustic regularities, simulating the introduction of accented speech by shifting correlations across acoustic speech dimensions. It offers the benefit of measuring the contribution of specific acoustic dimensions in signaling speech categories, and thereby allows for quantification of the degree adaptation that is necessary to test the magnitude and direction predictions described above.

In the paradigm, listeners' baseline perceptual weights are first assessed across a two-dimensional acoustic space. In the present studies, voice onset time (VOT) and fundamental frequency (F0) vary across *beer* and *pier* whereas spectral quality (SQ, related to the first and second formants' relationship) and duration (DU) vary across *set* and *sat.* When stimuli are sampled equiprobably from the acoustic space, the acoustic dimensions' pattern of influence on speech categorization is correlated in a manner that aligns with English speech regularities (e.g., higher F0s and longer VOTs each signal *pier*). Likewise, as is typical of speech categories, the two dimensions do not contribute equally in signaling speech category identity. For example, VOT carries a stronger *perceptual weight* than F0; it is a better indicator of /b/ - /p/ category membership (Abramson & Lisker, 1985; Francis et al., 2008; Haggard et al., 1970; Whalen et al., 1993). Equiprobable sampling across the acoustic space provides an estimate of listeners' baseline perceptual weights, which broadly align with patterns of native-language experience.

With this baseline established, the paradigm models an encounter with speech acoustics that deviates systematically from native-language norms, as in accented speech, by selectively sampling speech stimuli to manipulate short-term regularities. Although this manipulation is subtle, presented across a common voice, and largely unbeknownst to the listeners, it rapidly induces shifts in the perceptual weight with which the acoustic dimensions signal speech categories (Idemaru & Holt, 2011; Liu & Holt, 2015). Returning to the *beer-pier* example, English listeners rely primarily upon VOT to inform category identity, with F0 as a secondary signal. The influence of F0 on speech categorization can be observed, especially, in holding the strong VOT signal perceptually ambiguous. In this case, higher F0s reliably result in more *pier* responses whereas lower F0s result in more *beer* responses—a pattern that aligns with distributional norms of American English speech. Yet, upon encountering a block of trials that conveys an "accent" in which the F0 $\times$ VOT relationship is reversed from English norms, listeners rapidly down-weight reliance on F0 such that it no longer reliably signals /b/ versus /p/. These results, replicated many times and apparent for vowel as well as consonant categorization, demonstrate the flexibility of the mapping from acoustics to speech categories upon encountering short-term regularities that depart from the norm (Idemaru & Holt, 2011, 2014, 2020; Lehet & Holt, 2017, 2020; Liu & Holt, 2015; Schertz et al., 2016; Zhang et al., 2021; Zhang & Holt, 2018).

Idemaru and Holt (2011) proposed that the primary, heavily perceptually weighted acoustic dimension (e.g., VOT in the case of *beer-pier*; Idemaru & Holt, 2011, 2014, 2020; Lehet & Holt, 2020; Zhang & Holt, 2018) serves to disambiguate category identity in the context of the accent. Following the logic of the supervised learning account sketched above, this would generate predictions about the patterns of speech input typically associated with the category, including secondary acoustic dimensions like F0. Upon introduction of the accent, the relationship of the secondary dimension falls out of alignment with typical patterns of speech experience. The hypothesis is that this generates a mismatch signal that drives the observed dimension reweighting of the secondary, here F0, dimension.

As described above, this model makes two specific predictions: both the *direction* and the *magnitude* of the perceptual adaptation should be modulated by the effectiveness of category activation by the dominant dimension. With respect to the direction of adaptation effects, the expectation is that the acoustic dimension that best signals category membership (VOT in the example above) should

elicit category activation, which will guide perceptual reweighting of the secondary dimension (F0 in the example). Thus, manipulating listening contexts to influence which acoustic dimension is dominant would be expected to impact the direction of perceptual adaptation—that is, which acoustic dimension is perceptually reweighted.

The results of Schertz et al. (2016), which capitalized on differences across English and Korean, are intriguing with regard to the direction prediction. Whereas both languages rely on voice-onset-time (VOT) and fundamental frequency (F0) for /b/ - /p/ categorization, they do so somewhat differently. Most English listeners rely more on VOT than F0; VOT has greater perceptual weight. Korean listeners show much more variability in baseline perceptual weighting: some listeners consistently rely most on VOT for /b/ - /p/ categorization whereas other listeners primarily perceptually weight F0 or weight it approximately equally with VOT. Schertz et al. confirmed these distinctive patterns of baseline perceptual weights by examining perception across speech stimuli equiprobably sampling VOT and F0 acoustic dimensions. In line with the direction prediction sketched above, baseline perceptual weights impacted the pattern of reweighting observed upon introduction of an accent in the dimension-based statistical learning paradigm. When VOT was the primary dimension, introduction of short-term regularity that reversed the typical F0 × VOT relationship led to down-weighting of F0. In contrast, when F0 was the primary dimension the F0 × VOT reversal resulted in VOT down-weighting. Listeners who perceptually weighted VOT and F0 approximately equivalently did not evidence much down-weighting of either dimension. By a supervised learning account, the dominant, perceptually weighted acoustic dimension provides category activation and when the secondary dimension mismatches its category-typical norms, it is perceptually down-weighted.

The present study builds upon the quasi-experimental cross-linguistic design of Schertz et al. (2016) to causally manipulate *which* acoustic dimension carries greater perceptual weight across different listening contexts among the same sample of listeners. We introduce distinct listening contexts by capitalizing on two signal processing techniques that flip which of two acoustic dimensions best signals phonetic category identity—that is, which carries greatest perceptual weight. This allows us to manipulate baseline perceptual weights as a function of listening context, influence which acoustic dimension will best signal category membership and, correspondingly, which is predicted to be down weighted upon introduction of an accent.

The second prediction is that the *magnitude* of perceptual adaptation effects will be a function of category activation. We predict that the more effective the primary, perceptually weighted acoustic dimension in unambiguously activating a phonetic category, the greater the magnitude of perceptual adaptation—that is, the greater the down-weighting of the secondary dimension when short-term regularities of the accent conflict with language norms. To test this prediction, we examine the accuracy with which listeners categorize stimuli that are perceptually *unambiguous* across the primary dimension as a proxy for category activation. We use the difference in categorization of two stimuli differing solely on the secondary acoustic dimension as a measure of perceptual adaptation via down-weighting. We predict a relationship such that the more accurate listeners are in categorizing unambiguous stimuli according to the primary dimension, the greater the phonetic category activation, and the stronger the perceptual adaptation evidenced as down-weighting of the secondary dimension.

The present study examines the direction and magnitude predictions across two experiments, one focused on consonant categorization of *beer-pier* and the other on vowel categorization across *set-sat*. In each, we first assess baseline perceptual weights in two listening contexts designed to modulate listeners' reliance on a particular acoustic dimension for speech categorization. These same listeners also categorize speech in each listening context as short-term regularities in speech are manipulated to introduce an accent, a correlation across dimensions that is reversed from English norms. This manipulation, observed to produce down-weighting of the secondary dimension in prior studies (e.g., Idemaru & Holt, 2011; Liu & Holt, 2015), is expected to interact with the listening context and the baseline perceptual weights it produces. We expect both the direction and magnitude predictions to be conditioned on the effectiveness of our experimental manipulation of listening context at the level of individual listeners. Thus, we test the direction and magnitude predictions by examining correlations across baseline perceptual weights, accuracy of categorizing unambiguous stimuli, and perceptual weight across the secondary dimension as tests of the direction and magnitude predictions.

Experiment 1

As noted above, most native English listeners perceptually weight VOT more than F0 in /b/ - /p/ speech categorization, although both acoustic dimensions contribute to /b/ - /p/ identity (Abramson & Lisker, 1985). In Experiment 1, we introduce a signal manipulation known to influence baseline perceptual weights in /b/ - /p/ categorization (Holt et al., 2018; Winn et al., 2013): presentation of speech in noise shifts perceptual weights from VOT to F0, with listeners relying more on F0 to categorize speech-in-noise than clear speech. This presents an opportunity to have the same listeners categorize speech in contexts in which we expect greater reliance on F0 and VOT, respectively, to test the direction and magnitude predictions upon introduction of an accent with short-term regularities that violate English norms.

Method

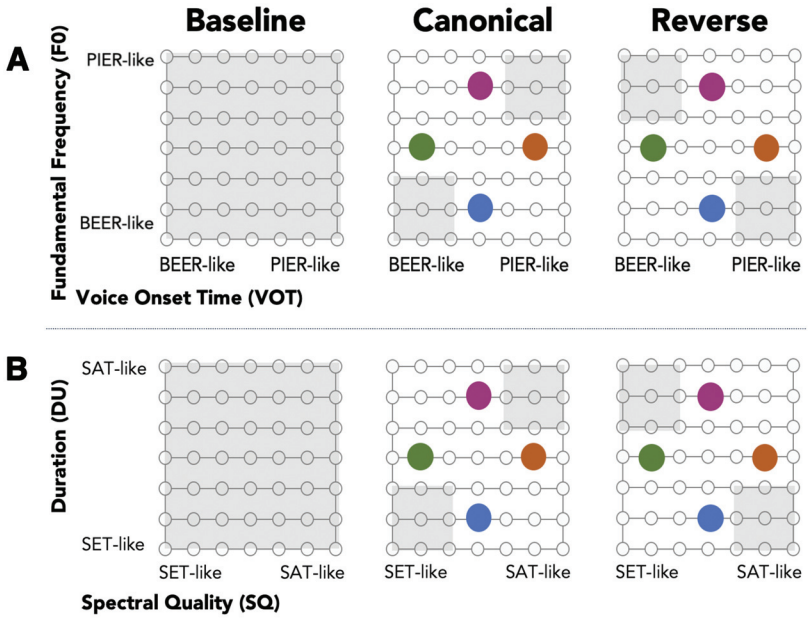
Participants

Sixty-one Carnegie Mellon University students (18–30 years old, average age 20 years; 38 women, 21 men, two preferred not to answer) received course credit or pay to participate in person. All reported normal hearing and English as the language used at home. Sample size was based on a power analysis informed by a pilot study. To detect a correlation of $r = .4$ at the significance level of .05 with a power of .8 requires at least 40 participants ('pwr' package in R; Cohen, 1988), which we exceeded in sampling due to our expectation of individual differences in baseline perceptual weights.

Stimuli

The stimuli were based on those of Idemaru and Holt (2011). The stimulus space was defined across VOT and F0 dimensions with seven steps along each dimension creating 49 unique stimuli in a two-dimensional acoustic space (open circles, Figure 1A).

Figure 1
Experiment 1 and Experiment 2 Stimulus Distributions



Note. Across panels, each open circle represents a unique stimulus. The gray highlighted area indicates exposure stimuli sampled for a particular task. The Baseline block samples stimuli equiprobably to estimate baseline perceptual weights. The Canonical block samples stimuli according to a dimension correlation that aligns with English whereas the Reverse block presents the opposite correlation as an "accent." The large colored circles indicate test stimuli, which are present across blocks and provide a measure of the perceptual weight of a single dimension as the other dimension is held constant and perceptually ambiguous. (A) Voice onset time (VOT) and fundamental frequency (F0) vary across beer-pier stimuli in Experiment 1. (B) Spectral quality (SQ) and duration (DU) vary across set-sat stimuli in Experiment 2. See the online article for the color version of this figure.

This acoustic space served as the basis for two stimulus sets: clear speech and speech-in-noise.

Clear speech tokens were created from natural productions of *beer* and *pier* by a female native-English speaker. Using Praat (Boersma, 2006), VOT was manipulated in 5-ms steps between 5 ms and 35 ms, creating a series varying perceptually from *beer* to *pier* (Abramson & Lisker, 1985). Vowel onset fundamental frequency (F0) of the following vowel was then measured from each step in the series and linearly interpolated into 20-Hz steps between 200 Hz to 320 Hz to create the 49-step stimulus grid, with RMS amplitude matched across the 49 stimuli.

Speech-in-noise tokens were generated from these clear speech tokens, in a manner similar to the approach described by Winn et al. (2013). Using Praat, speech-shaped noise was generated by extracting the long-term average spectrum from a 5-min NPR interview segment and filtering white noise with this spectrum. The noise was then RMS-matched to the same level as the clear speech and an onset and offset 200-ms cosine ramp was applied. The manipulated noise was then added to clear speech. The procedure was performed for all 49 clear speech tokens to create stimuli for the speech-in-noise condition.

Procedure

Seated in front of a computer monitor in a sound-attenuated booth, participants heard speech tokens diotically over headphones (Beyer DT-100) and responded whether the stimulus was *beer* (Z key) or *pier* (M key) on a standard keyboard. There was a one-second pause separating trials and no feedback. Visual prompts BEER and PIER aligned with the relative position of the response keys. Participants responded after the offset of the audio, guessing if they were unsure.

Baseline Block. All participants first categorized each of the 49 tokens six times in a randomly presented order (total of 294 trials, see left panel in Figure 1A) in speech-in-noise first, and then in clear speech (gray highlighted regions, Figure 1). These trials were separated into three 98-trial blocks for each listening condition.

Canonical/Reverse Blocks. Short-term input regularities were presented across 132 exposure trials reflecting canonical English patterns of F0 × VOT, or the reversed pattern (Figure 1A, middle and right panels). Exposure trials were subsampled from the same stimuli space used in baseline blocks, with sampling shown highlighted in gray in Figure 1A. In a Canonical block, listeners heard 18 exposure stimuli and four test stimuli (colored, filled markers in Figure 1A) six times each, with the 132 total trials presented in a random order. In the Canonical block, the sampling of exposure stimuli reflected long-term English norms: long VOT was associated with high F0 and short VOT was associated with low F0. The Reverse block presented an accent with

the F0 × VOT relationship reversed such that long VOT was associated with low F0 and short VOT was associated with high F0. Four test stimuli were identical across both Canonical and Reverse blocks. Two of the four test stimuli were distinguished by F0 (pink, blue markers; Figure 1A), with perceptually ambiguous VOT; the other two test stimuli (green, orange markers; Figure 1A) varied in VOT with ambiguous F0. Exposure and test stimuli were intermixed randomly within a block. The categorization across test stimulus pairs measured perceptual weighting across VOT or F0 dimensions in the Canonical and Reverse blocks.

Overall, participants experienced a baseline block, a Canonical and a Reverse block with each block separated by self-time breaks. All participants completed this first for speech-in-noise and immediately thereafter for clear speech so that the clear speech did not unduly influence perception of speech-in-noise. Participants were only instructed to identify sounds as *beer* or *pier* and informed that the noisy sounds might be challenging to identify. They received no feedback and were not informed of the sampling of stimuli.

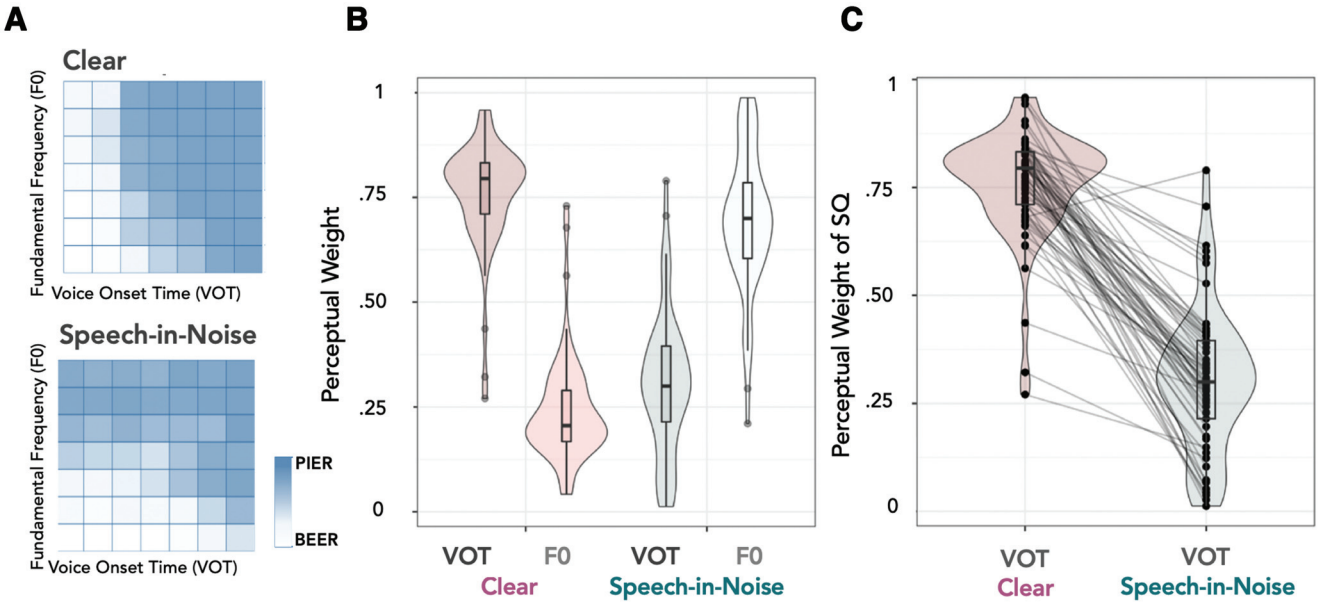
Results

Based on prior studies, we expected that VOT would be the dominant acoustic dimension signaling /b/ - /p/ categories in clear speech whereas F0 would be dominant in speech-in-noise (Holt et al., 2018; Winn et al., 2013). This prior work also led us to anticipate individual differences in the effectiveness of our listening context manipulation. Correspondingly, based on the hypothesis that category activation via the dominant dimension drives perceptual adaptation, we predicted that the *directionality* of adaptation playing out as dimension weighting on the secondary acoustic input dimension would vary across clear speech versus speech-in-noise contexts, and across listeners according to their susceptibility to the manipulation of which dimension dominates. Finally, we predicted that categorization accuracy of unambiguous exposure stimuli according to the dominant dimension (a proxy for distinctive category activation) would be associated with greater-*magnitude* down-weighting of the secondary dimension in the Reverse block.

Listening Context Shifts Baseline Perceptual Weights

We first examined baseline perceptual weights across clear and speech-in-noise listening contexts. Figure 2A shows the average proportion *beer* versus *pier* responses, illustrating that listeners use both VOT and F0 in categorization, and that the perceptual weight of dimensions varies across clear speech and speech-in-noise contexts. From these data we quantified perceptual weights using a regression model including VOT and F0 as predictors of category responses for each participant (Lehet & Holt, 2020; Liu & Holt, 2015). Coefficients were normalized to sum to one, as in prior research (Holt & Lotto, 2006). Figure 2B plots distributions of VOT and F0 perceptual weights as a function of listening context. The pattern aligns with expectations that American English listeners rely more on VOT (M = .76, 95% confidence interval, CI [.73, .79]) than F0 (M = .24, 95% CI [.21, .27]) in categorizing *beer* and *pier* in clear speech (Abramson & Lisker, 1985). Crucially, the presentation of speech-in-noise shifted reliance away from VOT (M = .30, 95% CI [.26, .34]) and toward F0

Figure 2
Experiment 1 Baseline Perceptual Weights



Note. (A) Heat maps of beer-pier consonant categorization across the voice onset time (VOT) and fundamental frequency (F0) acoustic input dimensions for clear speech (top) and speech-in-noise (bottom). Darker blue indicates more pier responses and lighter blue indicates more beer responses. (B) The data from (A) are summarized as violin and box plots of average normalized perceptual weights for VOT and F0 across clear speech and speech-in-noise. (C) The same data are plotted as violin and box plots for VOT perceptual weights to illustrate that almost all listeners (99.4%) relied less on VOT in noise than in clear speech. SQ = spectral quality. See the online article for the color version of this figure.

(M = .70, 95% CI [.66, .74]). Figure 2C replots the Figure 2B data to illustrate individual listeners' perceptual weight shift across listening contexts. A simple linear regression model with only condition (clear speech, speech-in-noise) and dimension (VOT, F0) as predictor variables (Akaike's information criterion, AIC = 61.99) and a full model were fit with listening context and dimension main effects as well as the Context × Dimension interaction term (AIC = −233.08). An F test revealed a significant Context × Dimension interaction (F = 570.88, $p < .001$, FDR-corrected). As illustrated in Figure 2C, nearly all participants (99.4%) exhibited this shift in perceptual weights across listening contexts, albeit to different degrees. Listening context strongly impacts which acoustic dimension is dominant in signaling category identity.

Baseline Perceptual Weights Predict Exposure Trial Categorization in the Reverse Block

The influence of listening context on the perceptual weight of acoustic dimensions was persistent across blocks. The dominant acoustic dimension at baseline robustly predicted categorization of exposure trials in the Reverse block. The perceptual weight of VOT in baseline categorization of clear speech was positively associated with Reverse block exposure trial categorization (r = .472, $p < .001$, FDR-corrected), indicative of unambiguous category activation by the dominant, VOT, dimension. Similarly, the F0 dimension—dominant in noise—predicted Reverse block exposure trial categorization for speech-in-noise (r = .583, $p < .001$, FDR-corrected). Thus, manipulation of listening context from clear speech to speech-in-noise has a persistent influence on which acoustic dimension best drives category activation when the accent is introduced in the Reverse block.

Category Activation According to the Primary Dimension Predicts the Direction and Magnitude of Perceptual Adaptation

This established a context in which to examine the direction prediction: activation of the category by the dominant dimension is predicted to result in down-weighting of the secondary dimension in speech categorization upon introduction of the accent in the Reverse block. Thus, in this within-participant design, we expected down-weighting of *F0 in clear speech* and of *VOT in noise*. Moreover, we predicted that the more effective the primary, perceptually weighted acoustic dimension in unambiguously activating a phonetic category, the greater the magnitude of adaptation —that is, the greater the down-weighting of the secondary dimension when short-term regularities of the accent conflict with language norms (Reverse block).

We quantified category activation as the Reverse block exposure trial accuracy according to the primary dimension (VOT in clear speech; F0 in noise; Figure 2B) and perceptual weight as the difference in *pier* categorization responses across test stimulus pairs varying along a dimension (with the opposing dimension held constant, see Figure 1). Figure 3 plots the relationship of category activation and the perceptual weight for the primary (Figure 3A and 3B) and secondary (Figure 3C and 3D) acoustic dimensions for clear speech (Figure 3A and 3C) and speech-in-noise (Figure 3B and 3D). First examining the dominant acoustic dimensions, we observe the expected relationship: listeners' reliance on the dominant dimension is positively associated with category activation for both clear speech (r = .645, $p < .0001$; FDR-corrected, Figure 3A) and speech-in-noise (r = .454, $p < .001$, FDR-corrected, Figure 3B). Those participants who more heavily weighted VOT in clear speech or F0 in speech-in-noise also were more accurate in Reverse block exposure trial categorization. This is another demonstration that the dominant acoustic dimension measured at baseline drives category activation of the exposure stimuli that convey the Reverse block accent.

We next examined the direction prediction, using the difference in test stimulus categorization of the secondary dimension as an index of perceptual adaptation (Figure 3C and 3D). Smaller differences in test stimulus categorization indicate greater the down-weighting of the dimension in response to the accent—that is, greater perceptual adaptation. Thus, we predict a negative association with category activation: the more effective category activation by the dominant dimension, the smaller categorization differences of test stimuli (i.e., the degree of down-weighting of the dimension). Indeed, category activation quantified as Reverse block exposure trial categorization accuracy according to the dominant dimension was negatively associated with the degree to which the secondary dimension differentiated categorization responses across test stimuli for both clear speech (r = −.308, p = .024, FDR-corrected, Figure 3C) and speech-in-noise (r = −.419, p = .002, FDR-corrected, Figure 3D). For each listening context, the greater the category activation by the dominant dimension, the less participants relied on the secondary dimension in signaling test stimulus category identity in the context of the accent that reversed its typical relationship to the dominant dimension. This is the down-weighting consistent with perceptual adaptation to the accent (Idemaru & Holt, 2011).

Note, too, that the acoustic dimension across which down-weighting is apparent varies according to the within-listener manipulation of listening context. When VOT is effective in signaling category identity in clear speech, listeners down-weight F0 in response to a reversal in the F0 × VOT regularity that conveys the accent (Figure 3C). When VOT is less effective in noise, F0 drives category activation and down-weighting is apparent in VOT (Figure 3D). Thus, the directionality of the effect—which acoustic dimension is impacted by the regularity conveying the accent—is predicted by category activation via the dominant dimension, which can be manipulated across listening contexts.

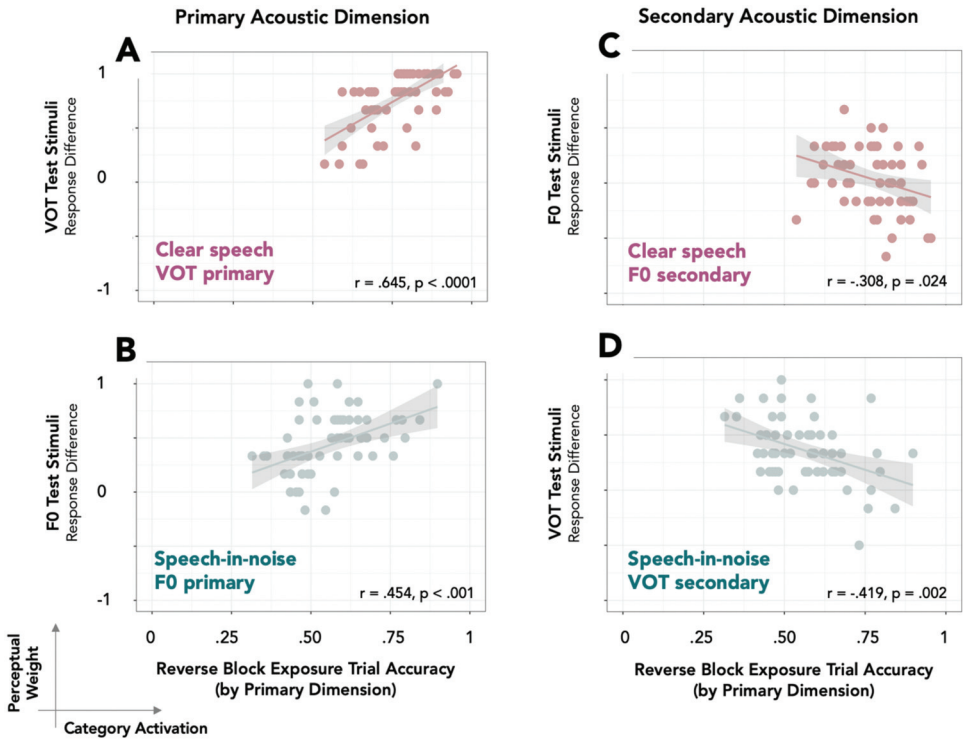
Finally, we examined the magnitude prediction. We expected that the more effective the primary, perceptually weighted acoustic dimension in unambiguously activating a phonetic category, the greater the magnitude of perceptual adaptation as evidenced by down-weighting of the secondary dimension. The linear relationships evident in Figure 3C and 3D support this prediction. The more effectively participants used the dominant dimension to signal category identity of Reverse block exposure stimuli, the greater the down-weighting of the secondary dimension. This is in line with an account that depends upon graded phonetic category activation to elicit perceptual adaptation when acoustic input mismatches category-specific expectations from long-term experience. The magnitude of perceptual adaptation (down-weighting) observed was related to successful category activation across stimuli that conveyed the accent.

Experiment 2

Following the logic of Experiment 1, Experiment 2 tests the directionality and magnitude predictions with an accent introduced

Figure 3

The Direction and Magnitude of Perceptual Adaptation are Predicted by the Dominant Acoustic Dimension, Experiment 1



Note. The top two panels (red, A and C) present data from clear speech. The bottom panels (blue, B and D) present the same participants' responses to speech-in-noise. Each plot shows the relationship of category activation, defined as the accuracy of exposure trial categorization in the Reverse block defined according to the primary dimension estimated at baseline on the *x*-axis. The *y*-axis plots corresponding perceptual weights for the primary (A and B) or secondary acoustic dimension (C and D). [FDR-corrected statistics]. See the online article for the color version of this figure.

to influence vowel categorization across *set* and *sat*, as described by Liu and Holt (2015). A group of listeners categorized speech sampled equiprobably across the two-dimensional spectral quality by duration (SQ × DU) acoustic space (baseline perceptual weights), and also across sampling that produced short-term regularities consistent with long-term English norms (Canonical) and violating those norms as an accent (Reverse), as shown in Figure 1B. The same listeners categorized both clear speech and vocoded speech, the latter of which is known to diminish the contribution of spectral dimensions like SQ while preserving temporal acoustic dimensions like DU (Shannon et al., 1995). Whereas Experiment 1 was conducted in person in a controlled laboratory setting with university student participants, Experiment 2 was conducted via online recruitment across a global sample of American English listeners tested using their own computer and headphones in their home environment.

Method

Participants

Seventy participants were recruited online using Prolific (prolific.co) and received payment for their time (18–35 years old, average age 27.5 years; 32 women, 37 men, one nonbinary). All reported normal hearing and American English as the native language used at home. Sample size was based on our power analysis from Experiment 1, with oversampling to assure a sampling of individual differences could be observed in this online study.

Stimuli

The stimulus space, based on that of Liu and Holt (2015), was defined across SQ and DU dimensions with seven steps along each dimension creating 49 unique stimuli in a two-dimensional acoustic space as shown in Figure 1B. This acoustic space served as the basis for two stimulus sets: clear and vocoded speech.

Clear speech tokens were created from natural productions of *set* and *sat* by a female native-English speaker. The first five formant trajectories were extracted in Praat (Boersma, 2006) across the steady-state portions of the vowels, spliced from their respective words, and interpolated in equal steps between /ɛ/ and /æ/, then resynthesized to create a seven-step spectral series. This created the SQ dimension, varying predominantly in first and second formant center frequencies. Vowel steady-state duration was manipulated to vary from 175 ms to 475 ms, creating the DU dimension. Each of the resulting vowels was concatenated with

the same /s/ and /t/ segments to create a 49-stimulus grid varying perceptually from *set* to *sat*.

The vocoded stimulus set was generated from these clear speech tokens, in a manner described previously (Hervais-Adelman et al., 2011; Shannon et al., 1995) using Praat (Boersma, 2006). The frequency spectrum was divided into four logarithmically spaced analysis bands between 50 and 5500 Hz and clear speech tokens were filtered by these analysis bands. The resulting envelopes were applied to band-pass-filtered noise in the same frequency ranges; thereby, reducing spectral resolution while preserving temporal information. Our expectation was that the spectral degradation introduced by vocoding would cause listeners to rely more upon DU than SQ in categorizing vocoded *set-sat* stimuli.

Procedure

Participants were recruited via online research platform prolific.co, with the requirement that they use the Google Chrome browser on a laptop or desktop with wired headphones. A headphone check requiring dichotic presentation to succeed in a simple task assured compliance (Milne et al., 2021). The Gorilla.sc platform (Anwyl-Irvine et al., 2020) presented acoustic stimuli in lossless *.flac format and participants responded whether each speech stimulus was *set* (F key) or *sat* (J key). There was a one-second pause separating trials and no feedback. Visual prompts SET and SAT aligned with the relative position of the response keys. Participants responded after the offset of the audio, without time pressure. They were instructed to take their best guess if they were unsure of the identity of the utterance. In all other ways, the procedure mirrored that of Experiment 1.

Participants first categorized each of the 49 noise-vocoded speech tokens 6 times to establish baseline perceptual weights. These trials were separated into three 98-trial blocks, between which participants took brief self-timed breaks. This was followed by one block of in which participants categorized each of the 18 canonical exposure stimuli and four test stimuli six times. Lastly, participants categorized each of the 18 reverse exposure stimuli and the same four test stimuli six times. Immediately thereafter, they completed the same procedure for the clear stimuli. The order of vocoded speech and clear speech blocks was not counterbalanced because exposure to clear speech can influence perception of vocoded speech (Hervais-Adelman et al., 2008). Participants were informed of the distinction between the clear sounds and "degraded" sounds that might be challenging to categorize. They were not given feedback. In all, there were 1,116 total trials in a single online session.

Results

Based on the results of Liu and Holt (2015), we expected that SQ would be the dominant acoustic dimension signaling /ɛ/ and /ae/ categories, as in *set* and *sat*, in clear speech. Correspondingly, we predicted that vocoding would have its impact on this dominant spectral dimension, leading DU to be dominant in driving *set-sat* categorization of vocoded speech. Following the logic of Experiment 1, if category activation via the dominant, heavily perceptually weighted acoustic dimension drives adaptation then we predict that the directionality of the effect would play out differently across clear and vocoded speech, and across listeners according to their susceptibility to adjust perceptual weights across listening conditions. With categorization accuracy of unambiguous exposure stimuli as a proxy for category activation we also predicted that accuracy would be associated with greater-magnitudes perceptual adaptation (down-weighting of the secondary dimension), as assessed by the difference in test stimulus categorization.

Listening Context Shifts Baseline Perceptual Weights

We first examined baseline perceptual weights across clear and vocoded listening contexts. Figure 4A shows the average proportion *set* versus *sat* responses, illustrating that listeners use both SQ and DU in /ɛ/ and /ae/ categorization, and that the perceptual weight of dimensions varies across clear speech and vocoded speech contexts. Using the same data plotted in Figure 4A, we next quantified dimension perceptual weights using a regression model including SQ and DU as predictors of category responses for each participant (Liu & Holt, 2015; Lehet & Holt, 2020). Coefficients were normalized to sum to one to reflect the relative perceptual weight of each dimension (Holt & Lotto, 2006). Figure 4B plots distributions of SQ and DU dimension perceptual weights. These results align with expectations that American English listeners rely more on SQ (M = .80, 95% CI [.77, .84]) than DU (M = .20, 95% CI [.16, .23]) in categorizing *set* and *sat* in clear speech (Liu & Holt, 2015).

Vocoded speech shifted reliance away from SQ (M = .27, 95% CI [.23, .32])) and toward DU (M = .73, 95% CI [.68, .77]). The Figure 4B data are replotted in Figure 4C to illustrate individuals' shift in SQ perceptual weight across listening contexts. A simple linear regression model with only listening context (clear, vocoded) and dimension (SQ, DU) as predictor variables (AIC = 161.98) and a full model were fit with context and dimension main effects as well as the Context × Dimension interaction term (AIC = −164.52). An F test revealed a significant Context × Dimension interaction (F = 616.13, $p < .0001$; FDR-corrected). As illustrated in Figure 4C, nearly all participants (98.57%) exhibited this shift in perceptual weights across conditions, albeit to different degrees.

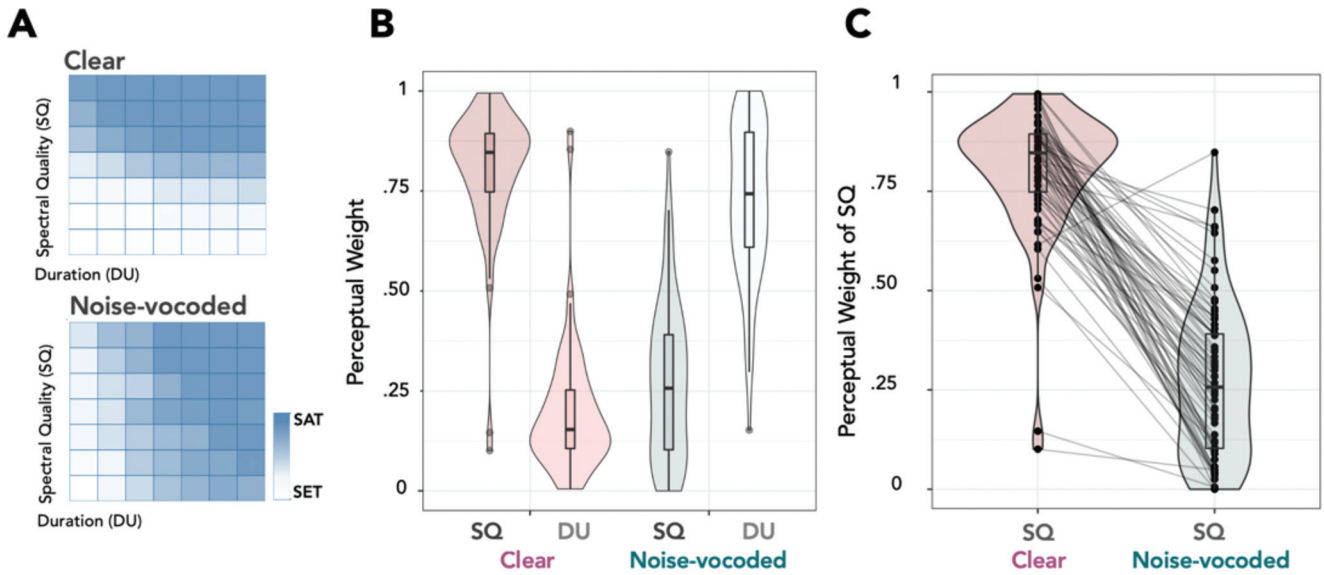
Baseline Perceptual Weights Predict Reverse Block Exposure Trial Accuracy

The influence of listening context on the perceptual weight of acoustic dimensions was persistent, as in Experiment 1. The dominant acoustic dimension at baseline robustly predicted exposure trial categorization in the Reverse block. The perceptual weight of SQ in baseline categorization of clear speech was positively associated with Reverse block categorization of exposure trials indicative of unambiguous category activation (r = .677, $p < .0001$; FDR-corrected). Similarly, the DU dimension—dominant in vocoded speech—predicted Reverse block exposure trial categorization (r = .481, $p < .0001$; FDR-corrected). Thus, listening context had a persistent influence on which acoustic dimension best signaled category for the Reverse block exposure trials conveying the accent.

Category Activation According to the Primary Dimension Predicts the Direction and Magnitude of Perceptual Adaptation

Relying on the same dependent variables as Experiment 1, we next examined the direction and magnitude predictions. As Figure 5 shows, category activation (the accuracy of Reverse block

Figure 4
Experiment 2 Baseline Perceptual Weights



Note. (A) Heat maps of vowel categorization across the spectral quality (SQ) and duration (DU) acoustic input dimensions for clear speech (top) and vocoded speech (bottom). Darker blue indicates more sat responses and lighter blue indicates more set responses. (B) The data from (A), presented as violin and box plots for average normalized perceptual weights across SQ and DU and for clear speech and vocoded speech. (C) The data from (A), plotted as violin and box plots for SQ weights only illustrate that almost all listeners (98.57%) relied less on SQ in vocoded speech compared with clear speech. See the online article for the color version of this figure.

exposure trials according to the dominant dimension) was related to the perceptual weight (the difference in *sat* categorization responses across test stimulus pairs) for the dominant (Figure 5A and 5B) and secondary (Figure 5C and 5D) acoustic dimensions for clear speech (Figure 5A and 5C) and vocoded speech (Figure 5B and 5D). These relationships differed across listening context and acoustic dimension in a manner that aligns with observations from Experiment 1.

First examining the primary acoustic dimensions, we find the expected relationship: category activation according to the dominant dimension is related to perceptual weight of that same dimension for both clear speech ($r = .8$, $p < .0001$; FDR-corrected, Figure 5A) and vocoded speech ($r = .676$, $p < .0001$; FDR-corrected, Figure 5B). Those participants who relied more heavily on SQ in clear speech or DU in vocoded speech also were more accurate in exposure trial categorization in the Reverse block in which the accent was conveyed. The dominant acoustic dimension measured at baseline drives category activation of the exposure stimuli as they convey the accent and listening context flips which dimension is dominant in the Reverse block.

We examined the direction prediction using the difference in test stimulus categorization of the secondary dimension as an index of perceptual adaptation. As in Experiment 1, category activation (Reverse block exposure trial accuracy according to the dominant dimension) predicted greater down-weighting for both clear speech ($r = -.248$, $p = .039$; FDR-corrected, Figure 5C) and vocoded speech ($r = -.455$, $p < .0001$; FDR-corrected, Figure 5D). The greater the effectiveness of category activation by the dominant dimension, the less participants relied on the secondary dimension in signaling test stimulus category identity in the context of the artificial accent. Said another way, greater category activation predicted greater perceptual adaptation. Listening context impacted the directionality of this relationship, flipping which acoustic dimension was dominant and which mismatched expectations, and was down-weighted. Thus, when SQ is effective in signaling category identity in clear speech, listeners down-weight DU in response to a reversal in short-term input regularities conveyed by the accent. When SQ is less effective in vocoded speech, DU drives category activation and down-weighting is apparent in SQ.

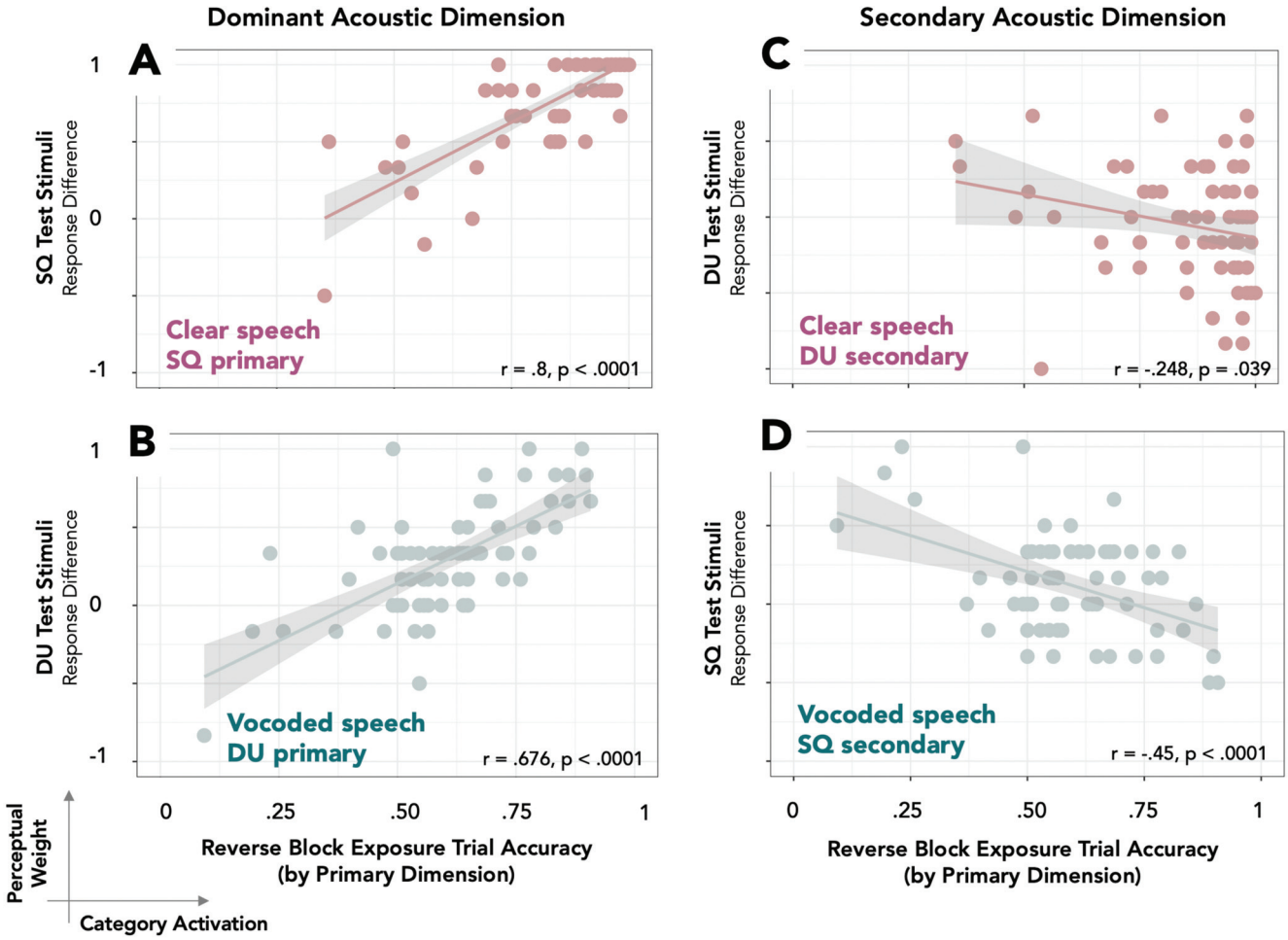
As in Experiment 1, we also observe evidence in support of the magnitude prediction. The linear relationships plotted in Figure 5C and 5D indicate that the more effectively participants use the dominant dimension to accurately signal category identity of Reverse block exposure stimuli, the greater the down-weighting of the secondary dimension. Thus, category activation predicts the magnitude of perceptual adaptation effects observed across listeners.

General Discussion

Speech perception presents a rich test-bed in which to examine how cognitive systems balance stability and flexibility. By adulthood, speech categorization reflects experience with the long-term regularities of the patterns of speech in one's native language such that some acoustic dimensions more robustly signal category membership than other dimensions. But, speech categorization remains flexible; listeners adapt to short-term regularities that depart from long-term norms. Understanding how the system flexibly adapts to short-term input regularities while not sacrificing long-term representations that have utility in most listening contexts is a central challenge for theories of speech perception, and

Figure 5

The Direction and Magnitude of Perceptual Adaptation are Predicted by the Dominant Acoustic Dimension, Experiment 2



Note. The top two panels (red, A and C) present data from clear speech. The bottom panels (blue, B and D) present data from the same participants' responses to vocoded speech. Each plot illustrates the relationship of category activation, defined as the accuracy of exposure trial categorization in the Reverse block defined according to the primary dimension estimated at baseline on the *x*-axis. The *y*-axis plots corresponding perceptual weights for the primary (A and B) or secondary acoustic dimension (C and D). (Statistics FDR-corrected.) SQ = spectral quality; DU = duration. See the online article for the color version of this figure.

contributes more broadly to understanding general learning mechanisms that might balance the tension of stability and flexibility in other domains. There is mounting evidence that listeners use a variety of disambiguating information sources to resolve ambiguity in short-term speech input and that this initiates rapid adjustments to speech categorization that persist for some time even after the disambiguating information is no longer available (Bertelson et al., 2003; Idemaru & Holt, 2011; Norris et al., 2003).

There is, as yet, no widespread agreement on the means by which the perceptual system rapidly adapts. Here, we investigated the implications of one proposal, supervised learning, which has been explored as a model (Guediche et al., 2014; Norris et al., 2003; Vroomen et al., 2007). Guediche et al. (2014, 2016), for example, proposed that when short-term speech input regularities that deviate from a listener's long-term speech representations are sufficiently disambiguated in the input to activate internal speech representations, expectations about the typical pattern of speech input related to that representation are available. A discrepancy between the expected and the actual sensory input may generate an error signal that drives adaptive changes in the mapping of acoustic input to speech representations. In this way, the disambiguating information (such as the primary acoustic dimension in the present studies) serves as a "teacher signal" inasmuch as it supports activation of the internal representation. To take a specific example in the context of the dimension-based statistical learning used in the present studies, Idemaru and Holt (2011) hypothesized that speech categories activated by a primary, heavily perceptually weighted, acoustic dimension could generate expectations about the typical mapping of secondary dimensions. When the short-term acoustic speech input violates these expectations, it generates an error signal that supervises adjustments to the mapping of speech to long-term representation, down-weighting the contribution of the secondary dimension to subsequent categorization. In this way, supervised learning accounts depend crucially on *activation of an internal speech representation*.

The present data offer the strongest data, to date, in support of this perspective. Specifically, we observe that adaptation to accent across both consonants and vowels exhibits patterns consistent with both the *directionality* and *magnitude* predictions of a supervised learning account. With respect to directionality, a supervised learning account predicts that the acoustic dimension that best signals category membership will elicit category activation that drives perceptual reweighting of the secondary dimension. When we causally manipulated which of two acoustic dimensions would be most likely to be dominant in a listening context, we found that listeners' patterns of adaptation shifted predictably to impact the nondominant, secondary dimension across both experiments.

Listeners were influenced by listening context. The acoustic dimension upon which they relied most heavily in consonant (Experiment 1) and vowel (Experiment 2) categorization shifted, even within a single experimental session and across the speech of a single talker. In turn, listening context had a robust influence on how the system adapted to a statistical regularity introduced in the Reverse block. The direction of influence was predicted by the *dominant* acoustic dimension, with perceptual adaptation playing out across the *secondary* dimension. We propose that the dominant dimension provides information with which to activate the phonetic category selectively, providing a prediction about how the secondary dimension tends to align with that representation. The mismatch introduced by the accent leads to adjustments in the effectiveness of the mismatched (secondary) dimension in subsequently signaling speech categorization. Moreover, at the level of individual participants, the degree to which this was true was predicted by the success of the dominant dimension in driving category activation, expressed as successful categorization responses. The current results demonstrate that the directionality of perceptual adaptation can be toggled by experimental manipulations of which acoustic dimension is dominant and which activates the category.

Each experiment also lends support for the magnitude prediction: the degree of the adjustment to the influence of the secondary dimension was linearly related to the strength of category activation by the dominant dimension. Listeners who were more successful at using the dominant acoustic dimension to categorize the exposure stimuli conveying the short-term regularity of the accent exhibited a greater degree of perceptual adaptation across the secondary dimension.

Our experimental results demonstrate that category activation via a dominant acoustic input dimension can drive dimension-based statistical learning across secondary acoustic dimensions, lending support for supervised learning accounts. To test the full explanatory power of supervised learning and its generality across different sources of disambiguating "teacher" signals and speech representations, future studies will be needed to investigate whether similar graded category activation effects exist in other paradigms, such as lexically or visually guided adaptation (Bertelson et al., 2003; Norris et al., 2003).

Two prior studies are especially interesting in this regard. In a dimension-based statistical learning paradigm like the one used here, Zhang et al. (2021) examined the impact of speech category activation on perceptual reweighting through top–down lexical knowledge. In their study, listeners identified nonminimal-pair word-nonword pairs, like *beef-peef* and *beace-peace*. Whereas, ordinarily, VOT would play a strong role in signaling the /b/ - /p/ category in such stimuli, Zhang and colleagues held it constant and perceptually ambiguous. Therefore, only lexical knowledge was available to disambiguate the input and potentially drive perceptual reweighting across the secondary, F0, dimension. They reasoned that if top–down feedback from lexical representations was sufficient to differentially activate phonetic categories according to the lexically consistent alternative (e.g., /b/ in *__eef* context but /p/ in *__eace* context), then it may be sufficient to drive dimension-based statistical learning—even when there was no unambiguous bottom–up acoustic input to differentiate the categories. Indeed, it was. In this way, these results align with the present research: activation of speech category representations, whether from bottom–up heavily-perceptually-weighted input dimensions or top–down lexical representations drives perceptual reweighting of the secondary input dimensions contributing to speech perception.

Another study reached somewhat similar conclusions using a very different paradigm. Guediche and colleagues (Guediche et al., 2016) presented listeners with monosyllabic English words severely distorted by vocoding and spectral shifting to create baseline conditions in which speech comprehension (indexed by open-set word recognition in which participants typed what they had heard) was extremely poor. Following classic research demonstrating adaptation in the mapping from speech acoustics to prelexical representations (Schwab et al., 1985), Guediche et al. presented listeners with written, orthographic information to disambiguate these distorted speech acoustics; the identity of the distorted spoken word was written on a screen after participants' responses. As in prior research, speech comprehension of the distorted acoustic input improved after just a short block of experience with the disambiguating orthographic information, and generalized to highly distorted words not experienced with the disambiguating information.

Next, Guediche et al. (2016) examined whether slightly less distorted speech, which might successfully activate lexical representations at least some of the time, might be sufficient to evoke the same effects as the disambiguating text on the screen. Their reasoning was similar to that described above: activation of an internal speech representation may evoke expectations about the typical mapping of acoustics, and drive adaptation to discrepant input. Indeed, when the supportive written, orthographic feedback was removed speech comprehension improved so long as the acoustic distortion was somewhat less severe and allowed for successful access to lexical representations (as measured by proportion correct word recognition). At posttest, listeners were better able to recognize distorted speech than they were before the exposure that supported successful lexical activation. In fact, reminiscent of the correlations reported in the present study, there was a significant positive relationship between word recognition accuracy during exposure to the less severe distortion and the magnitude of perceptual adaptation, measured as the improvement in word recognition before versus after exposure.

These prior studies are in accord with the present results and lend support for the claim that activation of an internal representation (here, phonetic category activation via a dominant input dimension) can drive perceptual adaptation. Future algorithmic modeling efforts will be needed to specify the detailed implementations of a supervised learning account, and whether they adequately capture the details of human behavior. For example, do

these effects evolve in a graded manner continuously across experience or is there a "trigger" point at which accumulated information about a shift in the short-term input statistics evokes adjustments? Tackling questions like these will advance a deeper understanding of putative mechanisms of perceptual adaptation, and whether this—or some other account—best accounts for extant data. The patterns of perceptual adaptation observed in the present studies indicate that any theoretical account will require an examination of a role for activation of internal speech representations in driving the effects.

References

Abramson, A., & 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. 25–33). Academic.

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods*, *52*(1), 388–407. https://doi.org/10.3758/s13428-019-01237-x

Bertelson, P., Vroomen, J., & De Gelder, B. (2003). Visual recalibration of auditory speech identification: A McGurk aftereffect. *Psychological Science*, *14*(6), 592–597. https://doi.org/10.1046/j.0956-7976.2003.psci_1470.x

Boersma, P. (2006). *Praat: Doing Phonetics by Computer*. www.praat.org

Bradlow, A. R., & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition*, *106*(2), 707–729. https://doi.org/10.1016/j.cognition.2007.04.005

Clarke, C. M., & Garrett, M. F. (2004). Rapid adaptation to foreign-accented English. *The Journal of the Acoustical Society of America*, *116*(6), 3647–3658. https://doi.org/10.1121/1.1815131

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Erlbaum.

Davis, M. H., Johnsrude, I. S., Hervais-Adelman, A., Taylor, K., & McGettigan, C. (2005). Lexical information drives perceptual learning of distorted speech: Evidence from the comprehension of noise-vocoded sentences. *Journal of Experimental Psychology: General*, *134*(2), 222–241. https://doi.org/10.1037/0096-3445.134.2.222

Francis, A. L., Kaganovich, N., & Driscoll-Huber, C. (2008). Cue-specific effects of categorization training on the relative weighting of acoustic cues to consonant voicing in English. *The Journal of the Acoustical Society of America*, *124*(2), 1234–1251. https://doi.org/10.1121/1.2945161

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

Guediche, S., Fiez, J. A., & Holt, L. L. (2016). Adaptive plasticity in speech perception: Effects of external information and internal predictions. *Journal of Experimental Psychology: Human Perception and Performance*, *42*(7), 1048–1059. https://doi.org/10.1037/xhp0000196

Haggard, M., Ambler, S., & Callow, M. (1970). Pitch as a voicing cue. *The Journal of the Acoustical Society of America*, *47*(2B), 613–617. https://doi.org/10.1121/1.1911936

Hervais-Adelman, A., Davis, M. H., Johnsrude, I. S., & Carlyon, R. P. (2008). Perceptual learning of noise vocoded words: Effects of feedback and lexicality. *Journal of Experimental Psychology: Human Perception and Performance*, *34*(2), 460–474. https://doi.org/10.1037/0096-1523.34.2.460

Hervais-Adelman, A. G., Davis, M. H., Johnsrude, I. S., Taylor, K. J., & Carlyon, R. P. (2011). Generalization of perceptual learning of vocoded speech. *Journal of Experimental Psychology: Human Perception and Performance*, *37*(1), 283–295. https://doi.org/10.1037/a0020772

Holt, L. L., Tierney, A. T., Guerra, G., Laffere, A., & Dick, F. (2018). Dimension-selective attention as a possible driver of dynamic, context-dependent re-weighting in speech processing. *Hearing Research*, *366*, 50–64. https://doi.org/10.1016/j.heares.2018.06.014

Holt, L. L., & Lotto, A. J. (2006). Cue weighting in auditory categorization: Implications for first and second language acquisition. *The Journal of the Acoustical Society of America*, *119*(5), 3059–3071. https://doi.org/10.1121/1.2188377

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

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

Idemaru, K., & Holt, L. L. (2020). Generalization of dimension-based statistical learning. *Attention, Perception, & Psychophysics*, *82*(4), 1744–1762. https://doi.org/10.3758/s13414-019-01956-5

Kraljic, T., & Samuel, A. G. (2005). Perceptual learning for speech: Is there a return to normal? *Cognitive Psychology*, *51*(2), 141–178. https://doi.org/10.1016/j.cogpsych.2005.05.001

Kraljic, T., & Samuel, A. G. (2006). Generalization in perceptual learning for speech. *Psychonomic Bulletin & Review*, *13*(2), 262–268. https://doi.org/10.3758/BF03193841

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

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

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

Milne, A. E., Bianco, R., Poole, K. C., Zhao, S., Oxenham, A. J., Billig, A. J., & Chait, M. (2021). An online headphone screening test based on dichotic pitch. *Behavior Research Methods*, *53*(4), 1551–1562. https://doi.org/10.3758/s13428-020-01514-0

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

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

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

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

Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science*, *270*(5234), 303–304. https://doi.org/10.1126/science.270.5234.303

Srinivasan, N. K., & Zahorik, P. (2013). Prior listening exposure to a reverberant room improves open-set intelligibility of high-variability sentences. *The Journal of the Acoustical Society of America*, *133*(1), EL33–EL39. https://doi.org/10.1121/1.4771978

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

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*(4), 2152–2159. https://doi.org/10.1121/1.406678

Winn, M. B., Chatterjee, M., & Idsardi, W. J. (2013). Roles of voice onset time and F0 in stop consonant voicing perception: Effects of masking noise and low-pass filtering. *Journal of Speech, Language, and Hearing Research*, *56*(4), 1097–1107. https://doi.org/10.1044/1092-4388(2012/12-0086)

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

Zhang, X., Wu, Y. C., & Holt, L. L. (2021). The learning signal in perceptual tuning of speech: Bottom up versus top-down information. *Cognitive Science*, *45*(3), e12947. https://doi.org/10.1111/cogs.12947

Received September 27, 2020
Revision received May 12, 2022
Accepted May 27, 2022 ■