Attention, Perception, & Psychophysics
https://doi.org/10.3758/s13414-019-01956-5

Generalization of dimension-based statistical learning

Kaori Idemaru [1] · Lori L. Holt [2]

Abstract
Recent research demonstrates that the relationship between an acoustic dimension and speech categories is not static. Rather, it is influenced by the evolving distribution of dimensional regularity experienced across time, and specific to experienced individual sounds. Three studies examine the nature of this perceptual, dimension-based statistical learning of artificially accented [b] and [p] speech categories in online word recognition by testing generalization of learning across contexts, and testing the effect of a larger word list across which learning is induced. The results indicate that whereas learning of accented [b] and [p] generalizes across contexts, generalization to contexts not experienced in the accent is weaker even for the same speech categories [b] and [p] spoken by the same speaker. The results support a rich model of speech representation that is sensitive to context-dependent variation in the way the acoustic dimensions are related to speech categories.

Keywords Speech perception · Statistical learning · Dimension-based learning · Cue weighting · Generalization

Introduction

Speech categories are characterized by multiple acoustic dimensions, some of which carry more information in signalling category affiliation than others (e.g., Abramson & Lisker, 1985; Idemaru & Guion, 2008; Lotto et al., 2004). By adulthood, perception tends to mirror the differential information carried by acoustic dimensions such that more diagnostic cues are given greater *perceptual weight;* they more effectively signal speech category membership. Adult listeners exhibit reliable perceptual weights that reflect regularities across acoustic dimensions experienced in the native language (Idemaru et al., 2012), and that are acquired across a long developmental course (Hazan & Barrett, 2000; Idemaru & Holt, 2013; Nittrouer, 1992). For example, although both voice-onset time (VOT) and the fundamental frequency (F0) of a following vowel signal voicing categories like [b] versus [p], VOT is a more robust signal of voicing category than F0; listeners give it more perceptual weight (Abramson & Lisker, 1985). Even more, listeners are sensitive to patterns of covariation across acoustic dimensions. English listeners have had long-term experience with the relationship between stop voicing categories and F0. Voiced stops, signalled by a shorter voice onset time (VOT), are typically produced with lower vowel F0, whereas voiceless stops, with a longer VOT, are usually produced with higher vowel F0. Thus, English listeners have long-term experience with this statistical pattern: shorter VOTs co-occur with lower F0s and longer VOTs with higher F0s. Correspondingly, adult speech categorization reflects this correlation. When the heavily perceptually weighted dimension, VOT, provides ambiguous information about category identity, listeners rely on F0 and do so in a manner that respects the VOT/F0 correlation experienced in long-term input. In this way, mature phonetic categorization reflects detailed regularities of long-term speech input.

Yet, the adult perceptual system remains flexible. In encountering short-term speech input that deviates from long-term regularities, such as in a foreign accent, perceptual weights rapidly adjust. Previous research (Idemaru & Holt, 2011, 2014; Lehet & Holt, 2017; Schertz, Cho, Lotto, & Warner, 2016; Zhang & Holt, 2018) has shown that when the long-term English relationship of F0 to voicing categories, as expressed in *beer* versus *pier* and also *deer* versus *tear*, reverses in local, short-term speech input creating an artificial "accent," listeners rapidly down-weight reliance on F0 in speech categorization. This *dimension-based statistical learning* has been demonstrated across spectral and durational

✉ Kaori Idemaru
idemaru@uoregon.edu

[1] Department of East Asian Languages and Literatures, University of Oregon, Eugene, OR 97403, USA

[2] Department of Psychology and Neuroscience Institute, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA, USA

Published online: 06 January 2020

acoustic dimensions signalling vowel categories as well (Liu & Holt, 2015) and influence listeners' own productions (Lehet & Holt, 2017).

Whereas dimension-based statistical learning occurs very rapidly, within ten trials of exposure to the short-term reversal of F0 and VOT in voicing (Idemaru & Holt, 2011), the pattern of learning does not necessarily lead to perceptual weights that squarely correspond to the short-term input statistics. Even after 5 days of exposure to an artificial accent that reverses the F0/VOT correlation speech input, the mapping of F0 to voicing category does not reverse (i.e., high F0s do not now signal voiced stops and low F0s voiceless stops; Idemaru & Holt, 2011). Instead, listeners down-weight F0 such that its influence in signalling voicing categories is reduced, or even eliminated.

This short-term adjustment of perception at the level of acoustic dimensions may be important for reliable speech perception in the face of extensive acoustic variability that exists in the signal. The previous work on dimension-based statistical learning contributes to a growing literature that has demonstrated that speech categorization is adjusted dynamically in response to a short-term deviation from the norm temporarily experienced in the speech signal. We now know that different types of information can guide such perceptual adjustment, including lexical information (Eisner & McQueen 2005; Kraljic & Samuel, 2006, 2007; Norris, McQueen, & Cutler, 2003; Reinisch & Holt, 2014), visual information (Bertelson, Vroomen, & de Gelder 2003; Reinisch, Wozny, Mitterer, & Holt, 2014; Reinisch & Holger, 2016; Vroomen, van Linden, de Gelder, & Bertelson, 2007;), phonotactic information (Culter, McQueen, Betterfield, & Norris, 2008), and statistical distributional information from the acoustics (Clayards et al., 2008; Idemaru & Holt, 2011, 2014; Liu & Holt, 2015; Zhang & Holt, 2018).

Subsequent research has sought to understand the generalization of perceptual learning to further determine the locus of its influence in speech processing. In this regard, the findings in the literature have not always been consistent, but more recent studies converge in suggesting that it is a fairly specific process. Studies on lexically guided perceptual learning have reported that whereas perceptual adjustment to an accented sound contrast (e.g., [s] vs. [ʃ]) encountered in various lexical items generalizes to a new lexical item, it does NOT generalize to a new talker (Kraljic & Samuel, 2007; McQueen, Cutler, & Norris, 2006), across manner of articulation (from [p]/[t] to [m]/[n]; Reinisch & Mitterer 2016), or position-conditioned allophonic variants (from word-final [ ]/[ɹ] to word initial [l]/[r]; Mitterer, Scharenborg, & McQueen, 2013). In some cases, learning was constrained to a specific vowel context: perceptual adjustment to [b]/[d] categorization did not generalize from [aba]/[ada] to [ibi]/[idi] or vice versa (Reinisch et al., 2014). There have been some studies that reported generalization in lexically guided perceptual learning, including generalization across places of articulation in stops (Kraljic & Samuel, 2005, 2006, 2007), across stops spoken by different talkers (Theodore et al., 2015), and position-conditioned allophonic variants (Mitterer, Cho, & Kim, 2016). Researchers explained that when generalization was obtained, it may have been due to acoustic similarities between adapter stimuli and test stimuli (Eisner & McQueen, 2005; Mitterer, Reinisch, & McQueen, 2018; Nusbaum & Morin, 1992; Reinisch & Holt, 2014).

Our own work has also suggested the highly specific nature of dimension-based statistical learning involving stop categorization (Idemaru & Holt, 2014). When listeners experienced a reversal of the relationship between VOT and F0 at one place of articulation (e.g., bilabial [b] vs. [p]) in short-term speech input, the perceptual weight of F0 was affected only for voicing categorization at this same place of articulation. Listeners continued to rely on F0 in voicing categorization at another place of articulation (e.g., alveolar [d] vs. [t]) even though all stimuli were produced by the same talker and the sounds occurred in the same phonetic context (i.e., [beer]/[pier] and [deer]/[tear]). Furthermore, when listeners experienced competing F0/VOT statistics across two places of articulation within the same block of trials, they showed evidence of tracking independent statistics at each place of articulation (see also Zhang & Holt, 2018). In experiencing a F0/VOT correlation reversal for [b] and [p] along with the canonical F0/VOT correlation for [d] and [t], the cumulative statistics across place of articulation nullify the F0/VOT correlation at each place such that there is no F0/VOT correlation in the short-term input. Yet, in this case, listeners down-weighted F0 in categorizing [b] and [p] while maintaining reliance on F0 in categorizing [d] and [t]. When the F0/VOT statistics in the input flipped in the course of experiment such that [d] and [t] were experienced with the reversed F0/VOT relationship while [b] and [p] followed the canonical English statistics, perception also shifted. In other words, dimension-based statistical learning does not seem to operate at the level of the abstract phonological feature stop voicing (at least within the approximately 40 min of exposure and testing), and instead, listeners appear to track separate (and opposing) distributional regularities across the pairs of contrasting sounds (i.e., [b]/[p] and [d]/[t]). This suggests that [b] and [d] are independent (Fig. 1a) instead of constituting a class (Fig. 1b) at the level at which dimension-based statistical learning operates. It is especially striking that this category-specific pattern of perception is observed across responses to speech stimuli produced by the same talker. These findings are consistent with the results of Maye and Gerken (2001), in which listeners learned to adjust stop categorization based on an exposure to short-term distribution of VOT ([t] vs. unaspirated [t]) deviating from the English norm at one place of articulation. The perceptual adjustment was evident only in categorization of sounds that deviated in short-term experience (e.g., alveolar stops: [t] vs.

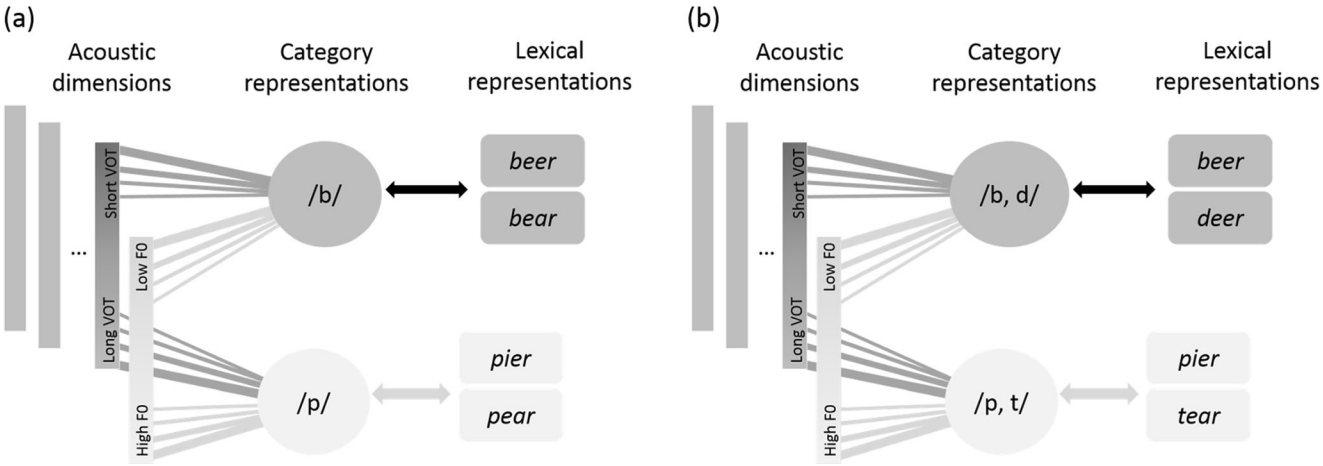

Fig. 1 Schematic illustration of the way acoustic dimensions and phonetic categories may be related for [b]/[p] and [d]/[t]. The width of the lines connecting acoustic dimensions to category representations indicates the activation strength, or perceptual weight, of the effectiveness of the dimension information in driving category activation. (**a**) The pattern in which [b]/[p] and [d]/[t] are separated; (**b**) the pattern in which the two voicing sounds are grouped together

unaspirated [t]) and not for another stop place (e.g., velar stops: [k] vs. unaspirated [k]), likewise indicating that learning did not occur at the level of an abstract phonological feature, stop voicing.

Prior work thus intriguingly suggests highly specific learning arising across distributions of acoustic dimensions that is resistant to phoneme generalization and supports tracking of independent regularities across speech categories, even for the same talker (Idemaru & Holt, 2014; Maye & Gerken, 2001; Zhang & Holt, 2018). The current study further examines generalization in dimension-based statistical learning. The results of Idemaru and Holt (2014) are consistent with learning that is reliant on activation of speech categories (e.g., [b] and [p]), and not phonological classes (e.g., [b], [p], [d] and [t]). However, the implication of Idemaru and Holt (2014) that the learning operates at the level of speech categories has not been fully tested. For example, this learning may be lexically specific. If lexical context constrains learning, perceptual adjustment of [b]/[p] voicing categorization elicited with short-term F0/VOT statistics experienced across *beer* and *pier*, for example, would not generalize to *bear* and *pear.* Such a constraint would suggest that the way speech categories, [b] for example, are related to acoustic dimensions is specific to the lexical context in which speech categories appear. If the learning does operate at the level of phonetic category regardless of such context, perceptual adjustment of [b]/[p] voicing categorization learned through an experience with *beer* and *pier* would generalize to *bear* and *pear.* Such generalization would suggest that speech categories, [b] for example, are related to acoustic dimensions in a uniform manner whether they appear in *beer* or *bear* (Fig. 1a).

Dimension-based statistical learning of vowels indeed appears to operate at the phonetic category level: perceptual adjustment of [ɛ]/[æ] categorization learned through short-term exposure to an accent in *set* and *sat* generalized to *setch* and *satch* (Liu & Holt, 2015), demonstrating that learning occurred with [ɛ]/[æ] categorization in general (regardless of the [s_t] context or the [s_ tʃ] context). The results of Liu and Holt (2015), however, suggest potentially graded generalization. When generalization was tested from the *set – sat* pair to the *setch – satch* pair that shared acoustically identical [sɛ] and [sæ] (with the *setch – satch* pair being created with the [sɛ] and the [sæ] portions extracted from the *set – sat* pair), generalization was robust. When generalization was tested from the female-produced *set – sat* pair to another *setch – satch* pair that were modified using "change gender" function on Praat (Boesma & Weenink, 2017), resulting in male-like speech, generalization was attenuated. Whereas the presence of generalization does demonstrate that the locus of perceptual adjustment is at the phonetic category level (i.e., [ɛ] and [æ]) rather than the lexical level (i.e., specific to *set* and *sat*), the acoustic similarity of speech seems to influence generalization. Moreover, the lexical contexts tested were highly similar, precluding a stringent test of the wider scope of generalization.

Yet, many open questions remain. For example, does dimension-based statistical learning across consonants operate at the phonetic category level and what factors affect generalization of this learning? Whereas there is evidence of this learning operating at the category level for vowels (Liu & Holt, 2015), it is yet to be tested for consonants. In Experiment 1, listeners experienced an introduction of an artificial accent that reversed the F0/VOT relationship in one word frame (e.g., *beer* and *pier*), and we examined generalization of perceptual adjustment of [b]/[p] categorization to the same voice producing [b]/[p] in a different word frame (*bear* and *pear*) for which vowel varied but the context was otherwise highly similar. This manipulation provides a means of

investigating whether the learning is specific to the word frame, i.e. context, or general to speech categories regardless of frame. We use the term context in this report to refer to the linguistic frame in which [b] and [p] were produced. We use this term in a general sense, not specifically referring to the lexical context, because it is as yet unknown whether it is the lexical context, the diphone context, the vowel context, or yet another kind of context (e.g., acoustics) that may influence generalization of learning.

To foreshadow, the Experiment 1 manipulations yield weak generalization. Thus, in Experiment 2, we tested learning with a new set of test stimuli to verify the results were not specific to the stimuli used in Experiment 1. In Experiment 2, listeners experienced an artificial accent in one word frame (*beer* and *pier* or *bear* and *pear*), and we examined generalization of perceptual adjustment to [b]/[p] categorization in a new word frame (*bill* and *pill*). In Experiment 3, listeners experienced accent across multiple word frames with a diversity of contexts (*beer, pier, bill, pill, best, pest* and one non-lexical pair *borth* and *porth*) and we examined generalization of perceptual adjustment to the same voice with a different context (*bear* and *pear*).

Experiment 1

Experiment 1 examined the extent to which dimension-based statistical learning generalizes to a context not experienced in the artificial accent. One group of listeners experienced an artificial accent reversing the F0/VOT correlation experienced across [b] and [p] in *beer* and *pier*. The impact of this accent was assessed across [b] and [p] categorization in *beer* and *pier* (experienced in the accent), as well as *bear* and *pear* (generalization context, not experienced in the accent). Another group of listeners experienced the artificial accent across *bear* and *pear* utterances, with the impact of this exposure tested across both *beer* and *pier* (generalization context), as well as *bear* and *pear* (experienced context). The two lexical pairs were selected for their highly similar contexts and vowel similarity.

Method

Participants A total of 62 native English listeners with normal hearing participated for credit or a small payment. Thirty-two participants were exposed to the F0/VOT reversal in *beer-pier* stimuli (Beer group) and 30 participants were exposed to the F0/VOT reversal in *bear-pear* stimuli (Bear group). Participants were either university students or employees. None of them participated in other experiments reported here.

Stimuli Natural utterances of *beer, pier, bear,* and *pear* were digitally recorded (22.05 kHz) in a sound-attenuated booth by an adult female native English speaker. The pairs of end-points were selected for similar duration (385 ms) and F0 contour. Stimuli were then constructed using progressive cross-splicing of the end-point tokens so they vary perceptually from [b] to [p] along the VOT series (McMurray & Aslin, 2005). As the first step to create the VOT continuum, 15 splice points were identified at the onset of each of the voiced and voiceless end-point tokens, with steps of approximately 2- or 3-ms increments and always at zero crossings. As a next step, the first interval of the voiced token was removed, starting at the onset and ending at the first splicing point (2–3 ms from the onset). A corresponding interval (2–3 ms from the onset) from the voiceless token was extracted and inserted at the beginning of the voiced token that was now missing the onset. This created a new end-point token. Then, another interval of the voiced token was removed, starting at the onset and ending at the second splicing point (4–6 ms from the onset). A corresponding interval (4–6 ms from the onset) from the voiceless token was extracted and inserted at the beginning of the voiced token missing the 4–6 ms at the onset. These steps were repeated to replace all 15 intervals. From the resulting sounds, those with VOT values of 0, 10, 15, 20, 25, 30, 40, and 50 ms were retained as stimuli. Sounds with -10 ms VOT were created by taking 10 ms of pre-voicing in the voiced production of the same speaker and inserting it before the burst of the voiced endpoint token (VOT = 0 ms).

The F0 contour of the two VOT series (*beer-pier* and *bear-pear*) was then manipulated such that the onset fundamental frequency of the vowel was adjusted to vary from 170 Hz to 190 Hz (Low F0s), and from 240 Hz to 260 Hz (High F0s) in three 10-Hz steps. For each sound, the F0 contour of the original production was manually manipulated using Praat 5.3 (Boersma & Weenink, 2017) to adjust to the target-onset F0 values. From the onset, the F0 decreased quadratically to 150 Hz at the end of the word. The high and low values of F0 and the contour modelled the natural production of the speaker. The stimuli were then normalized to the same root-mean-square amplitude (75 dB).

Design and procedure

Baseline categorization task Listeners first categorized *beer-pier* and *bear-pear* VOT series to measure the baseline influence of F0 on voicing judgments. Stimuli varying along VOT in nine steps (-10, 0, 10, 15, 20, 25, 30, 40, and 50 ms) and along F0 at two levels (180 Hz and 250 Hz) were presented in random order ten times each, blocked for *beer-pier* and *bear-pear* with the block order counter-balanced across participants.

Participants were seated in front of a computer monitor in a sound booth. Each trial consisted of a spoken word presented diotically over headphones (Beyer DT-150) and visual display of words, *beer* and *pier*, or *bear* and *pear*, corresponding to

the two response choices each with a designated key number presented on a monitor. The experiment was delivered under the control of E-prime experiment software (Psychology Software Tools, Inc.). Participants were instructed to press the key corresponding to the word they heard as quickly as possible across a total of 360 trials in the categorization task (9 VOTs × 2 F0s × *beer-pier, bear-pear* × 10 repetitions[1]). This provided a test to confirm that participants' voicing judgments reflected experience with the long-term F0/VOT correlation typical of English, as has been observed in many previous studies (Abramson & Lisker, 1985; Haggard et al., 1970; Idemaru & Holt, 2011, 2014; Whalen et al., 1993).

Word recognition task Immediately following the baseline test, the word recognition task exposed listeners to canonical, reversed and canonical F0/VOT correlations via exposure stimuli and monitored reliance upon F0 in categorizing the test stimuli that possessed the value of VOT that was in the center of the VOT continuum.

As shown in Fig. 2, exposure stimuli had perceptually *unambiguous* VOT values signaling the voicing categories. However, the relationship between the VOT and F0 changed across the course of the experiment, exposing listeners to a short-term deviation in the F0/VOT correlation typical of English voicing categories (Abramson & Lisker, 1985). These exposure stimuli served as "teaching signal" as their VOT unambiguously signaled category affiliation, [b] or [p], while at the same time providing statistical information regarding how F0 was mapped to the categories.

Test stimuli, in contrast, had perceptually more *ambiguous* VOT values (20 ms). F0 exerts the strongest influence on voicing perception when VOT is ambiguous (Abramson & Lisker, 1985; Idemaru & Holt, 2011, 2014), and thus the VOT-neutral test stimuli provide an opportunity to observe subtle changes in listeners' use of the F0 as a function of experienced changes in the correlation between F0 and VOT. In Fig. 2, gray cells indicate exposure stimuli, whereas black cells were test stimuli. Test stimuli were interspersed among the exposure stimuli throughout the experiment. Test stimuli allow us to assess the difference in voiceless responses across High F0 and Low F0 as an estimate of reliance on F0 in voicing categorization.

For Beer group listeners, exposure stimuli (gray, Fig. 2) only included *beer* and *pier*, whereas the test stimuli (black) consisted of both *beer/pier* and *bear/pear* VOT-neutral tokens. For Bear group listeners, exposure (gray) stimuli only included *bear* and *pear*, whereas the test stimuli (black) consisted of both *beer/pier* and *bear/pear* VOT-neutral tokens. "*Beer-pier*" (and "*bear-pear*") with a hyphen is used in this report to refer to the continuum of stimuli, or categorization of stimuli, that spanned between the exemplar (endpoint) *beer* and the exemplar *pier*. "*Beer/pier*" (and "*bear/pear*") with a slash is used to refer to the VOT-neutral, category-ambiguous stimuli.

As in previous studies (Idemaru & Holt, 2011, 2014), participants were exposed to the shift of F0/VOT correlation from the canonical English pattern (high F0 with voiceless stops and low F0 with voiced stops) (e.g., Abramson & Lisker, 1985) to the reversed pattern and then back to the canonical English pattern in a continuous word-recognition task. In more detail, in the first block, the Beer group listeners heard *beer* and *pier* exposure words with the familiar canonical English F0/VOT correlation: *beer* had lower F0s, whereas *pier* had higher F0s on the vowel. This was the same for the Bear group listeners except that exposure words were *bear* and *pear*. In these canonical-correlation exposure stimuli, three perceptually *unambiguous* short VOT values (-10, 0, and 10 ms, heard as [b]) were combined with three low F0s (170, 180, and 190 Hz), whereas three long VOT values (30, 40, and 50 ms, heard as [p]) were combined with three high F0s (240, 250, and 260 Hz). In the second block, the F0/VOT correlation in the exposure words was reversed such that listeners heard [b] and [p] with an F0/VOT correlation opposite their long-term experience with English (reverse correlation). Note that for exposure stimuli, VOT always unambiguously signaled the voicing category. Although F0 was correlated with VOT, it was never essential for speech categorization, which could be accomplished entirely with the unambiguous, and perceptually most heavily weighted, VOT. In the last block, the F0/VOT correlation in the exposure stimuli returned to the canonical English F0/VOT correlation.

The exposure stimuli (five each of *beer* and *pier* for the Beer group, and five each of *bear* and *pear* for the Bear group) were presented in 20 random orders for a total of 200 exposure trials per block to expose listeners to the canonical or reversed F0/VOT correlation. The *beer/pier* and *bear/pear* test stimuli were constant in each block. The four test stimuli (*beer/pier, bear/pear* × 2 F0s) were presented ten times for a total of 40 test trials in each block in a random order, interspersed among the exposure stimuli. The test stimuli were not described to participants, and they were not differentiated from exposure stimuli by task or instructions. In the entire experiment, there were 600 exposure trials and 120 test trials, consistent with Idemaru and Holt (2011, 2014).

The procedure and apparatus for this task were identical to those for the baseline categorization task. Trials proceeded continuously across the three blocks as listeners performed the word-recognition task. The block structure was implicit: participants were not informed that

[1] Due to technical issues, one level of stimuli, the middle of the continuum with 20 ms VOT, was presented 20 times instead of ten times for all tasks, except for *bear-pear* categorization by Bear group. This may have made *bear-pear* stimuli slightly more ambiguous for Bear group if anything at all; however, as VOT and F0 showed expected robust effects for both groups and pairs, we proceeded with the analysis.

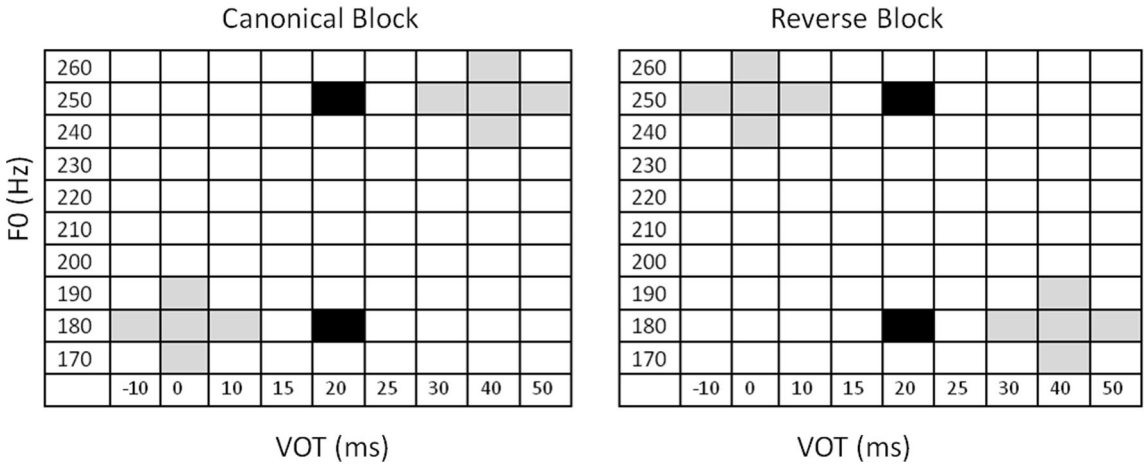

Fig. 2 Schematic illustration of stimulus sampling in the Canonical blocks (**left**) and the Reverse block (**right**) as a function of VOT and F0. Gray cells indicate exposure stimuli, and black cells indicate test stimuli

the experiment was divided into separate blocks, or that the nature of the acoustic cues would vary. The number of trials is consistent across this experiment and subsequent experiments (600 exposure and 120 test trials), and the entire session was completed in approximately 50 min.

Analysis

Baseline categorization All analyses presented in this study were performed using mixed-effect logistic regression (Breslow & Clayton, 1993; Jaeger, 2008) as implemented in the lme4 package (Bates, Mächler, Bolker, & Walker, 2014) in the R environment (R Core Team, 2016). The models analyzing listeners' categorization for [b] and [p] prior to exposure to F0/VOT correlation included VOT (continuous factor, centered), F0 (categorical factor, sum coded with Low F0 as the reference[2]), Group (categorical factor, sum coded with Bear group as the reference), stimulus pair ("Pair" henceforth, categorical factor, sum coded with *beer-pier* as the reference), and their interactions as fixed effects, as shown in Equation (1). The dependent variable was binary, indicating the listeners' response of [b] or [p] ([p] response = 1, and [b] response = 0). The model included a random intercept for Listener, but a random intercept for Group was not included as its variance was small and inclusion of the factor resulted in overfitting. This was the case for all other analyses reported in this study. Random slopes for VOT, F0, Pair, and their interactions over Listener were included as by-listener variation in the responses to these variables were expected or possible (e.g., Idemaru, Holt, & Seltman, 2012; Kong & Edwards, 2016; Lehet & Holt, 2017).

$$\text{Response}\sim\text{VOT}*\text{F0}*\text{Group}*\text{Pair} + (1 + \text{VOT}*\text{F0}*\text{Pair} \mid \text{Listener}) \quad (1)$$

Word-recognition test Mixed-effect logistic regression models were also used to analyze listeners' responses ([b] or [p]) to VOT-neutral test stimuli during exposure to the changing F0/VOT correlation. The model for this analysis included F0 (categorical factor, sum coded with Low F0 as the reference), Block (categorical factor, treatment coded with Reverse as the reference), Generalization Condition ("Condition" henceforth, categorical factor, treatment coded with Generalization as the reference), Group (categorical factor, sum coded with Bear group as the reference), and their interactions as the fixed effects, as shown in Equation (2). A random intercept for Listener was included. Random slopes for F0, Block, Condition, and their interactions were included, as by-listener variation in response to these variables was possible. We uncorrelated random factors to aide convergence problems.[3]

In this model, the coefficient for the sum-coded F0 reflects the differences in voiceless response between High F0 and Low F0; in other words, it corresponds to the weight of F0 in voicing categorization. Thus, an interaction effect between F0 and Block corresponds to changes in the weight of F0 to voicing categorization as a function of shifting F0/VOT correlation. With the reference level of Block set as Reverse, an F0 × Block interaction indicates a change in F0 weight from Canonical 1 to Reverse block, and from Reverse to Canonical 2. A significant F0 × Block interaction, therefore, is taken as evidence of learning in this study. With the reference level of

[2] The term reference is used to refer to the level coded as 0 in treatment coding schemes and the level coded as -1 in sum coding schemes.

[3] R does not correctly interpret random slope terms in an uncorrelated random effects formula if they are factors, as in the case of Equation (2). This is the case even when factors are contrast coded. We converted F0, Block, and Condition into numeric variables, and used them in the formula.

Condition set as Generalization, an F0 × Block interaction, in fact, corresponds to the change of F0 weight across blocks in the Generalization frame, a condition critical to our research question. The data and analysis code are available at https://osf.io/wx7py/.

$$\text{Response} \sim \text{F0}*\text{Block}*\text{Condition}*\text{Group} + \left(1 + \text{F0}*\text{Block}*\text{Condition} \parallel \text{Listener}\right) \quad (2)$$

Our prior work has repeatedly shown that the perceptual weight of the F0 dimension in voicing categorization decreases in Reverse block (Idemaru & Holt, 2011, 2014; Zhang & Holt, 2018). But, we have always tested perceptual adjustment of F0 using VOT-neutral test stimuli (e.g., VOT-neutral *beer/pier*) that aligned in the word frame with exposure stimuli (accented [b] and [p] in *beer* and *pier*). The present research question was, therefore, whether this online perceptual adjustment of F0 on voicing categorization generalizes to a new context.

Results

Baseline categorization Figure 3 illustrates proportion of voiceless responses across the nine-step VOT dimension and High F0 and Low F0 in categorizing *beer-pier* and *bear-pear* stimuli by Beer group and Bear group. At this point, exposure is balanced; the exposure introducing short-term deviations in the VOT/F0 correlation is only relevant in the next task. The critical aspect of this analysis was to verify that both groups of listeners relied on F0, in addition to VOT, at baseline prior to the later exposure test.

The results in the form of regression tables are provided in the Appendix. Only the results relevant to our research question (i.e., main effects of VOT and F0, and interactions involving F0) are interpreted here. We found significant effects of VOT and F0, as expected (VOT: $\hat{\beta}$ = 0.28, SE = 0.09, z = 20.57, p < 0.01; F0: $\hat{\beta}$ = 0.89, SE = 0.06, z = 15.98, p < 0.01). As indicated by the coefficients and also seen in Fig. 3, there were more voiceless responses for longer VOT values and for High F0. We also found several significant interactions involving F0, indicating that the magnitude of F0 effect varied due to other variables: the F0 effect was slightly stronger for shorter VOT values (VOT*F0: $\hat{\beta}$ = -0.03, SE = 0.01, z = -4.03, p < 0.01), and this F0 × VOT interaction slightly varied across groups (VOT*F0*Group: $\hat{\beta}$ = -0.02, SE = 0.01, z = -2.73, p = 0.01). There was also a trend that the pattern was varied slightly across *bear-pear* categorization and *beer-pier* categorization (VOT*F0*Group*Pair: $\hat{\beta}$ = -0.01, SE = 0.01, z = -1.89, p = 0.06).

These results verified that F0 indeed influenced [b]/[p] categorization in both the *bear-pear* frame and the *beer-pier* frame for both groups of listeners. The observed effect of F0 at baseline reflects listeners' long-term experience with lower F0s associated with voiced categories and higher F0s associated with voiceless categories. The baseline F0 was approximately equivalent across the two groups prior to the exposure test.

Categorization of exposure stimuli In the main experiment, listeners responded to *exposure stimuli* that either conveyed the Canonical F0/VOT correlation, or an artificial accent with the Reverse F0/VOT correlation. Across all of these stimuli, VOT unambiguously signaled voicing category. We first examined listeners' responses to *exposure stimuli* with

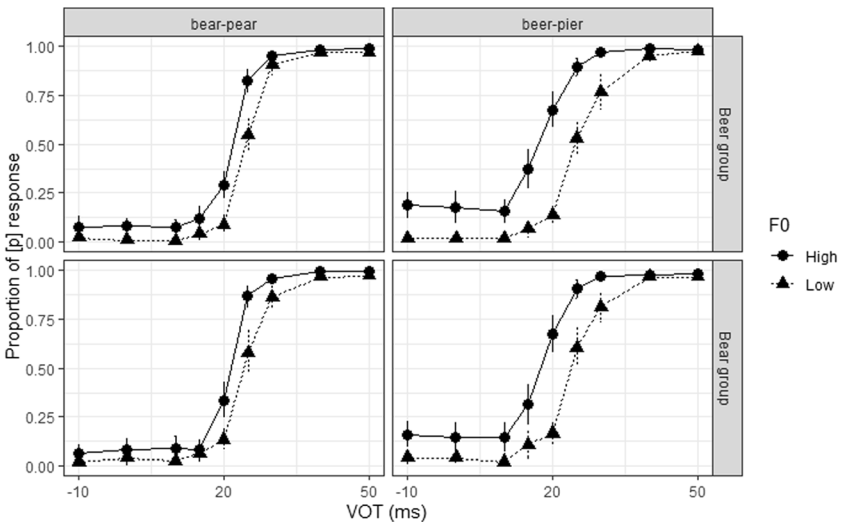

Fig. 3 Proportion of voiceless responses across nine steps of VOT (ms) and two levels of F0 in *bear-pear* categorization (**left**) and *beer-pier* categorization (**right**) by Beer group (**top**) and Bear group (**bottom**). We checked the two groups showed the effect of F0 at baseline prior to different exposure. Error bars indicate the 95% confidence interval of the mean

unambiguous VOTs (0 ms for [b] and 40 ms for [p]) to verify that listeners used VOT in [b]/[p] categorization. Note that these VOT values were the center values and the most frequent in the distribution of voiced and voiceless VOTs experienced across exposure stimuli (Fig. 2). The mean proportion of expected responses collapsed for *beer* and *bear* (voiced) and *pier* and *pear* (voiceless) was high: voiced, M = 0.91, SD = 0.28; voiceless, M = 0.95, SD = 0.20, indicating that listeners indeed used VOT as expected for voicing categorization.

Categorization of test stimuli The primary concern of the current analysis was to determine the extent with which the influence of F0 signalling voicing categories decreased in the Reverse block compared to other blocks (indicated by a significant F0 × VOT interaction), and whether this learning generalized to voicing categorization in a new, Generalization word frame (at the reference level of Condition). We interpret all effects that are statistically significant at $p < 0.05$.

The results of the regression model are presented in Regression Table 2 (Appendix), and Fig. 4 illustrates predicted probability of [p] responses and standard error of these predictions across blocks separately for Beer group and Bear group. Robust changes in the influence of F0 was evident across blocks in the Experienced frame (Fig. 4). As seen in Regression Table 2, we obtained a trend of significant four-way F0*Block*Condition*Group interaction, which is interpreted below. The critical F0 × Block interaction was significant between Canonical 1 and Reverse block, but not between Reverse and Canonical 2 block, when the level of Condition was Generalization (F0*Canonical1: $\hat{\beta} = 0.47$, SE = 0.12, z = 3.79, $p < 0.01$; F0*Canonical2: $\hat{\beta} = 0.15$, SE = 0.12, z = 1.26, p = 0.21). The significant F0 × Block × Condition interaction indicated that the F0 × Block interaction effects were greater (indicated by the positive sign on the coefficient) for the Experienced condition (F0*Canonical1*Experienced: $\hat{\beta} = 1.35$, SE = 0.17, z = 7.92, $p < 0.01$; F0*Canonical2*Experienced: $\hat{\beta} = 1.51$, SE = 0.16, z = 9.39, $p < 0.01$), which is evident in Fig. 4. The four-way interaction with Group indicated that there was a trend that this effect was stronger for Beer group from Reverse to Canonical 2 block (F0*Canonical2*Experience*Beer: $\hat{\beta} = 0.31$, SE = 0.16, z = -1.96, p = 0.049).

These results demonstrated that whereas changes in reliance on F0 in voicing categorization across the blocks was robust in the Experienced frame, it was weaker and present only from Canonical 1 to Reverse blocks in the Generalization frame. In other words, perceptual down-weighting of F0 in voicing categorization generalized only weakly across *beer-pier* and *bear-pear* categorization.

In this experiment, listeners experienced a shift of the F0/VOT correlation characterizing [b] and [p] categories – a sort of "artificial accent." This exposure unambiguously signalled these consonants' category affiliation because the dominant VOT dimension unambiguously mapped to the categories in the manner typical of English. As such, it served as a "teaching" signal for how to map F0 to the [b]/[p] categories (Idemaru & Holt, 2011; Liu and Holt, 2015). When the input statistics reversed the mapping, deviating from a long-term English pattern, F0 was no longer effective in signalling voicing categories in word frames in which the artificial accent (mapping reversal) was experienced. This finding is consistent with prior research (Idemaru & Holt, 2011, 2014; Schertz et al., 2016; Zhang & Holt, 2018). Our current results present evidence that attenuation of the effectiveness of F0 signalling

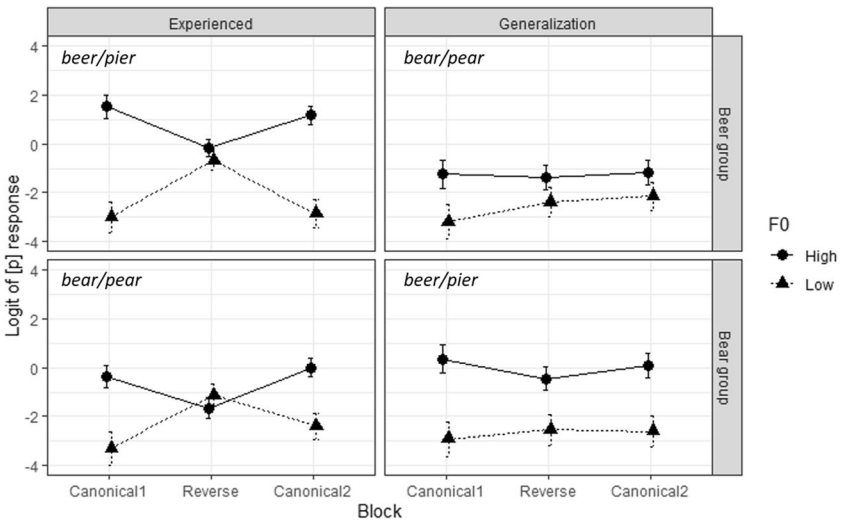

Fig. 4 Logit of [p] responses for High F0 (round marker) and Low F0 stimuli (triangle marker) across F0/VOT blocks for Experienced (left) and Generalization (right) conditions for Beer group (top) and Bear group (bottom). Error bars indicate the 95% confidence interval of the mean predicted by the model

voicing was also present, albeit weakly, in new word frames not experienced in the accent. In other words, perceptual adjustment of F0 dimension in voicing categorization did not robustly transfer from the word frame across which the accent was experienced to another word frame, but the specificity of the word frame did not completely prohibit generalization either. These results indicate that dimension-based statistical learning can occur at a pre-lexical level affecting perceptual processes contributing to phonetic categorization at a general level. Nevertheless, this adjustment was substantially reduced (from Canonical 1 to Reverse block) or absent (from Reverse to Canonical 2) when word frames were not experienced in the accent. It is important to note that this graded generalization occurred across categorization of the same sounds, [b] and [p], produced by the same talker. This means that whereas F0 was ineffective at signalling voicing categorization in one frame, it maintained somewhat greater effectiveness in another. The critical difference was whether or not the sound categories were experienced in the same frame in the accent. Recall that VOT and F0 values were identical across *beer-pier* stimuli and *bear-pear* stimuli. The results suggest that listeners do not "treat" the same talker's [b] and [p], and their identical VOT and F0, in the same manner. This is intriguing and it may suggest that this learning is robustly influenced by experienced regularities beyond the critical dimensions (i.e., F0 and VOT in this case). However, there is a possibility that these results were due to some characteristics specific to the particular pairs of words we tested (*beer-pier* and *bear-pear*). In order to examine whether a weak generalization is specific to this paring, we tested generalization of this learning in another word frame.

Experiment 2

Experiment 1 showed that dimension-based statistical leaning did not robustly generalize across *beer-pier* and *bear-pear* categorization. Experiment 2 tested whether a weak pattern of generalization observed in Experiment 1 can be observed in a different word frame. In Experiment 2, we used *bear-pear* and *beer-pier* as exposure stimuli, in which listeners experienced an artificial accent, consistent with Experiment 1, but we tested perceptual adjustment of F0 in a new word frame *bill/pill* not used in Experiment1. One group of listeners experienced the accent with shifting F0/VOT distribution across *beer* and *pier,* with the influence of this accent evaluated across [b] and [p] categorization in *beer* and *pier* (experienced in the accent) and *bill* and *pill* (generalization frame, not experienced in the accent). Another group of listeners experienced the accent across *bear* and *pear*, and were tested across *bear* and *pier* (experienced frame) and *bill* and *pill* (generalization frame).

Method

Participants A total of 62 native English listeners with normal hearing participated for credit or a small payment. They were either university students or employees. None of the listeners participated in other experiments reported here. Thirty-one participants experienced the reverse F0/VOT accent in *beer* and *pier* and were tested for generalization with *bill/pill* (Beer group) and another 31 participants experienced the reverse F0/VOT accent in *bear* and *pear* and were tested for generalization with *bill/pill* (Bear group).

Stimuli and procedure Stimulus construction methods and the procedure were the same as Experiment 1. Listeners in Beer group completed a baseline categorization task with *beer-pier* and *bill-pill* stimuli, and listeners in Bear group completed the baseline task with *bear-pear* and *bill-pill* stimuli. The two sets of stimuli were blocked for each group of listeners, and listeners completed a total of 360 trials (9 VOTs × 2 F0s × 2 continua × 10 repetitions[4]). In the word-recognition test, listeners in Beer group experienced exposure stimuli (gray cells in Fig. 2) *beer* and *pier,* and were tested with VOT-neutral test stimuli (black cells) *beer/pier* (Experienced condition) and *bill/pill* (Generalization condition). Listeners in Bear group experienced exposure stimuli *bear* and *pear,* and were tested with VOT-neutral test stimuli *bear/pear* (Experienced condition) and *bill/pill* (Generalization condition). As in Experiment 1, the word-recognition test proceeded from Canonical 1, to Reverse, and to the Canonical 2 blocks, with the block structure implicit to the participants, and with each block comprised of 200 exposure and 40 test trials. Throughout the experiment, participants' task was simply to identify each initial consonant as [b] or [p].

Analysis and results

Baseline categorization Figure 5 illustrates the results of baseline categorization. Responses were analysed using a regression model with VOT, F0, Group, Stimulus Pair ("Pair"), and VOT × F0 × Group and VOT × F0 × Pair interactions as the fixed effects, as shown in Equation (3). Group and Pair were not crossed in the model since stimulus pairs were not balanced across groups. The reference level for the categorical predictors were Low F0 (sum coded), Bear group (sum coded), and *bill/pill* (sum coded, with contrast 1 comparing the

[4] Due to technical issues, one level of stimuli, the middle of the continuum with 20 ms VOT, were presented 20 times instead of ten times for Bear group. This may have made the task slightly more ambiguous for Bear group if anything at all; however, as the factor Group did not make any significant effects, we proceeded with the analysis.

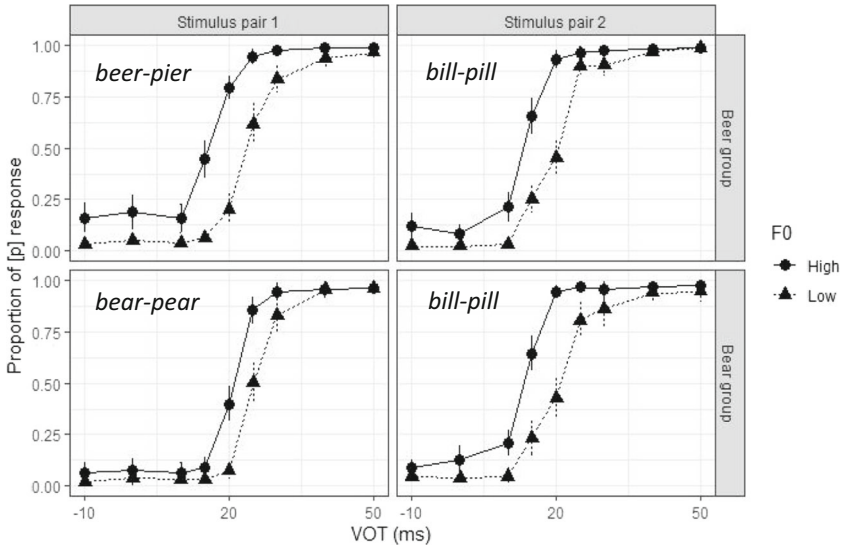

Fig. 5 Proportion of voiceless responses across nine steps of VOT (ms) and two levels of F0 for *beer-pier* and *bill-pill* series by listeners in Beer group (top), and *bear-pear* and *bill-pill* series by listeners in Bear group (bottom)

reference level and *bear-pear*). Random intercept for Listener, and by-listener random slopes for VOT × F0 × Pair were included. Random effects factors were uncorrelated to address overfitting.

$$\text{Response} \sim \text{VOT}*\text{F0}*\text{Group} + \text{VOT}*\text{F0}*\text{Pair} + \left(1 + \text{VOT}*\text{F0}*\text{Pair} \parallel \text{Listener}\right) \quad (3)$$

The results are reported in Regression Table 3 (Appendix). We found significant effects of VOT and F0 (VOT: $\hat{\beta} = 0.28$, SE = 0.02, z = 16.28, $p < 0.01$; F0: $\hat{\beta} = 1.04$, SE = 0.06, z = 16.27, $p < 0.01$). As expected, there were more voiceless responses for longer VOT values and for High F0. VOT and F0 did not show a significant interaction with Group, indicating that the two groups did not differ in their reliance on VOT and F0 for voicing categorization. Significant interaction involving F0 indicated that there was modulation of F0 effects due to VOT and Pair: F0 effect was slightly stronger for shorter VOT values (VOT*F0: $\hat{\beta} = -0.02$, SE = 0.01, z = -3.26, $p < 0.01$), weaker for *bear-pear* than *bill-pill* (F0* bear/pear: $\hat{\beta} = -0.21$, SE = 0.07, z = -2.86, $p < 0.01$) and stronger for *beer-pier* than *bill-pill* (F0* beer/pier: $\hat{\beta} = 0.27$, SE = 0.07, z = 3.78, $p < 0.01$). These results confirmed that F0 influenced [b]/[p] categorization in all word frames tested here, and that the two listener groups did not differ from each other in the voicing categorization.

Categorization of exposure stimuli Listeners' responses to *exposure stimuli* with unambiguous VOTs (0 ms for [b] and 40 ms for [p]) showed high proportion of expected responses: M = 0.91, SD = 0.29 for voiced; M = 0.98, SD = 0.15 voiceless for *beer* and *pier*, and M = 0.93, SD = 0.25 for voiced; M = 0.95, SD = 0.22 for voiceless for *bear* and *pear*, confirming that listeners indeed used the perceptually unambiguous VOT appropriately for categorization.

Categorization of test stimuli Responses were analyzed with F0, Block, Generalization condition ("Condition"), and Group as fixed factors in a regression model shown in Equation (4). The reference level of categorical factors were Low F0 (sum coded), Reverse block (treatment coded), Generalization condition (treatment coded), and Bear group (sum coded). The random-effects structure included random intercept for Listener, and random slopes for F0, Block, and Condition over Listener. Interaction terms were not included, since they showed no variance and models including them indicated issues overfitting. We compared the model fit between Equation (4) and a model with fully crossed random slopes (F0*Block*Condition). The comparison indicated that full model, while suffering an issue of overfitting, fit the data better ($\chi 2(4) = 32.725$, $p < 0.01$). However, since the general pattern of results were consistent across the two sets of results, we report the results of the model in Equation (4). The corresponding results from the full model are provided in the footnote.

$$\text{Response} \sim \text{F0}*\text{Block}*\text{Condition}*\text{Group} + \left(1 + \text{F0} + \text{Block} + \text{Condition} \parallel \text{Listener}\right) \quad (4)$$

The results are presented in Regression Table 4 (Appendix) and Fig. 6 illustrates predicted probability of [p] responses and standard error of these predictions across blocks separately for

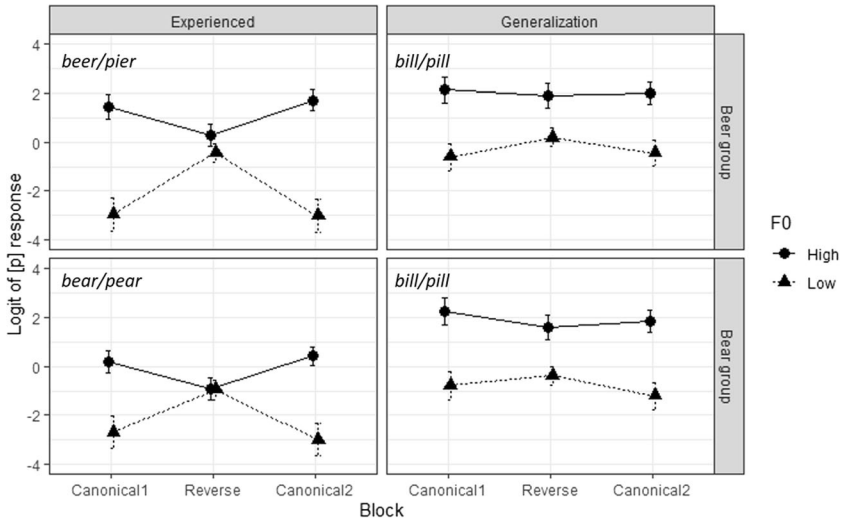

Fig. 6 Logit of [p] responses for High F0 (round marker) and Low F0 stimuli (triangle marker) across F0/VOT blocks in the Experienced condition (**left**) and the Generalization condition (**right**) by Beer group (**top**) and Bear group (**bottom**). Error bars indicate the 95% confidence interval of the mean predicted by the model

Beer group and Bear group. As in Experiment 1, robust changes in the influence of F0 across blocks was evident for the Experienced condition (Fig. 6). First of all, the four-way F0 × Block × Condition × Group interaction was not significant, indicating the response pattern was consistent across Bear group and Beer group. More importantly, the F0 × Block interaction was significant (F0*Canonical1: $\hat{\beta} = 0.58$, SE = 0.12, z = 4.94, $p < 0.01$; F0*Canonical2: $\hat{\beta} = 0.52$, SE = 0.11, z = 4.54, $p < 0.01$[5]), indicating that reliance on F0 in voicing changed across blocks when the level of Condition was Generalization. And this effect interacted with Condition such that the effect was stronger, as expected, in the Experienced condition (F0*Canonical1*Experienced: $\hat{\beta} = 0.86$, SE = 0.16, z = 5.43, $p < 0.01$; F0*Canonical2*Experienced: $\hat{\beta} = 1.17$, SE = 0.16, z = 7.32, $p < 0.01$[6]). These results indicated that down-weighting of F0 occurred in both Experienced and Generalization frames, but the magnitude of this effect was smaller in the Generalization condition.

In this experiment, we tested generalization of perceptual adjustment of F0 in [b]/[p] categorization in a new word frame, *bill/pill*, to determine whether the results we obtained in Experiment 1 could be replicated. Indeed they could. Effectiveness of F0 as voicing cue was attenuated in a new word frame unexperienced in the accent, but the attenuation was much weaker compared to attenuation of F0 influence in voicing categorization in word frames that were heard in the accent. These results suggest that generalization of dimension-based statistical learning is very conservative and may be quite specific to experienced regularities of the relevant speech categories in the input signal.

Experiment 3: Effect of multiple words

In the two preceding experiments, we observed robust dimension-based statistical learning of stop voicing in the experienced word frame. Listeners rapidly attenuate reliance on the F0 dimension in experiencing an artificial accent that reverses the typical English F0/VOT correlation. This F0 down-weighting generalizes to other words distinguished by stop voicing contrast. F0 down-weighting, however, is graded; it is robust in word frames experienced in the artificial accent, but attenuated in word frames unexperienced in the accent, even when the stimuli are spoken by the same talker and possessing identical characteristics in the critical acoustic dimensions.

Experiment 3 examined a factor that may impact generalization. Experiments 1 and 2 exposed listeners to the artificial accent in a single word pair (e.g., *beer* and *pier*). In Experiment 3, we expanded the range of acoustic variation in the word frames in which the reversal of F0/VOT correlation is experienced by increasing the number of exposure words. Thus, in this case the accent is attested across a wider range of acoustic and word frame variability.

Whereas the input that listeners experienced in Experiments 1 and 2 did not necessarily indicate that the accent occurs specifically to the lexical items *beer* and *pier* (or to the specific experienced tokens of [b] and [p]), the input also does not necessarily indicate positively that the accent broadly impacts [b]s and [p]s across other contexts. Experiment 3 tests whether experiencing the artificial accent across *multiple* word frames impacts generalization. In Experiment 3, four word frames

[5] Per the full model: F0*Canonical1: $\hat{\beta} = 0.61$, SE = 0.12, z = 5.07, $p < 0.01$; F0*Canonical2: $\hat{\beta} = 0.54$, SE = 0.12, z = 4.68, $p < 0.01$.

[6] Per the full model: F0*Canonical1*Experienced: $\hat{\beta} = 0.86$, SE = 0.16, z = 5.35, $p < 0.01$; F0*Canonical2 Experienced: $\hat{\beta} = 1.15$, SE = 0.16, z = 7.14, $p < 0.01$.

comprised the exposure stimuli, with three lexical pairs, *beer, pier, bill*, *pill, best, pest* and one non-lexical pair *borth* and *porth*. We included the nonword frame so that the input would convey that this artificial accent is productive and occurs even across words that listeners have never heard before.

Method

Participants Thirty-two native-English listeners with normal hearing participated for credit or a small payment. They were either university students or employees. None of the participants participated in other experiments reported here. These 32 participants experienced the short-term F0/VOT reversal across three word frames and one non-lexical word frame.

Stimuli and procedure The stimulus construction methods and procedure were the same as Experiment 1. The baseline categorization task included *beer-pier* and *bear-pear* stimuli. The two sets of stimuli were blocked, and listeners completed a total of 400 trials (8 VOTs × 2 F0s × 2 continua × 10 repetitions + 1 VOT × 2 F0s × 2 continua × 20 repetitions[7]). The exposure stimuli (gray cells in Fig. 2) were eight words: *beer, pier, bill, pill, best, pest,* and non-lexical *borth* and *porth*. The test stimuli (black cells in Fig. 2) were *beer/pier* (Experienced condition) and *bear/pear* (Generalization condition), with neutral VOT and High and Low F0. Five unique tokens of each exposure word were each presented five times per block for a total of 200 exposure trials per block (8 words × 5 times × 5 tokens). The VOT-neutral test stimuli were each presented ten times per block for a total of 40 test trials (*beer/pier, bear/pear* × 2 F0s × 10 times). The total number of exposure trials (600) and test trials (120) was consistent with the previous experiments in this study.

Analysis and results

Baseline categorization Figure 7 illustrates proportion of voiceless responses categorizing *beer-pier* and *bear-pear* stimuli. Responses were analysed using a regression model with VOT, F0, Stimulus Pair ("Pair"), and VOT × F0 × Pair interaction as the fixed effects, as shown in Equation (5). The reference level for the categorical predictors were Low F0 and *beer-pier*. Random intercepts for Listener, and by-listener random slopes for VOT × F0 were included. Random slopes for Pair, and random intercepts for Group were not included as their variance was small and resulted in overfitting.

$$\text{Response}\sim\text{VOT}*\text{F0}*\text{Pair} + (1 + \text{VOT}*\text{F0} \mid \text{Listener}) \quad (5)$$

The results are reported in Regression Table 5 (Appendix). We found expected significant effects of VOT and F0 (VOT: $\hat{\beta}$ = 0.28, SE = 0.02, z = 11.60, p < 0.01; F0: $\hat{\beta}$ = 0.80, SE = 0.06, z = 12.28, p < 0.01), confirming that VOT and F0 affected voicing categorization. Significant VOT × F0 and F0 × Pair interactions indicated that the effect of F0 was greater for shorter VOT values (VOT*F0: $\hat{\beta}$ = -0.03, SE = 0.01, z = =3.60, p < 0.01), and smaller for *bear-pear* (F0*Pair: $\hat{\beta}$ = -0.03, SE = 0.01, z = 3.60, p < 0.01). These results confirmed that listeners used F0 to [b]/[p] categorization prior to the exposure experiment.

Categorization of exposure stimuli Listeners showed high rate of expected voicing categorization: voiced, M = 0.94, SD = 0.24 collapsed for *beer, best, bill,* and *borth*; voiceless, M = 0.94, SD = 0.24 collapsed for *pier, pest, pill* and *porth.* These results confirmed that listeners used VOT as expected in categorizing exposure stimuli conveying the F0/VOT correlation across blocks.

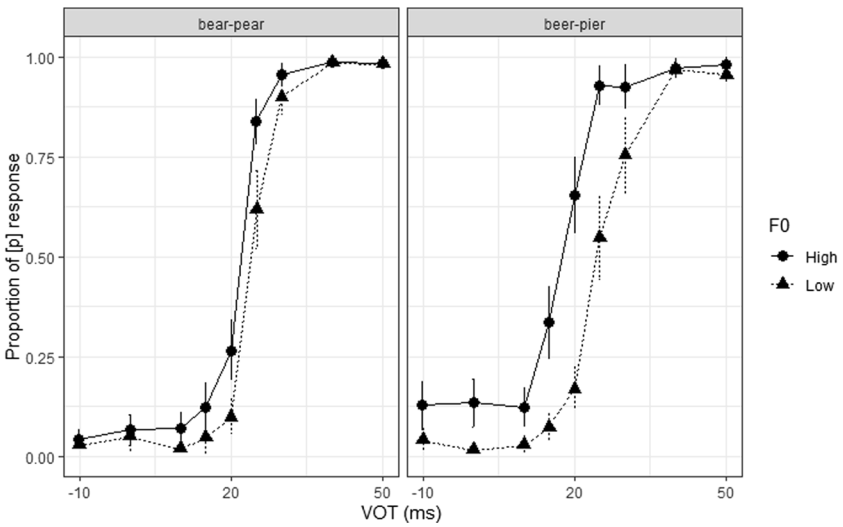

Fig. 7 Proportion of voiceless responses across nine steps of VOT (ms) and two levels of F0 for *bear-pear* (**left**) and *beer-pier* (**right**)

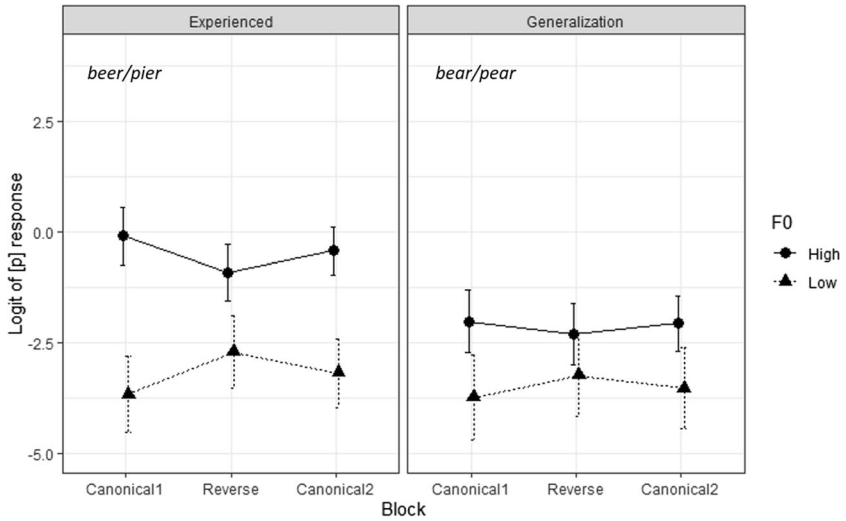

Fig. 8 Logit of [p] responses for High F0 (round marker) and Low F0 stimuli (triangle marker) across F0/VOT blocks for the Experienced condition (**left**) and Generalization condition (**right**). Error bars indicate the 95% confidence interval of the mean predicted by the model

Categorization of test stimuli Responses were analyzed with F0, Block, Generalization condition ("Condition") and their interactions as fixed effects in a regression model shown in Equation (6). The reference level of categorical factors were Low F0 (sum coded), Reverse block (treatment coded), and Generalization condition (treatment coded). Random intercept for Listener, and by-Listener random slopes for F0, Block, and Condition were included. The model including a random slopes for interaction terms over Listener indicated overfitting problems potentially due to correlations among the factors. The issue was not resolved by uncorrelating the random factors, but it was resolved by removing interaction terms. A comparison between Equation (6) and a model with fully crossed random slopes (F0*Block*Condition) indicated that there was no difference between the two models in model fit ($\chi 2(63) = 45.79$, $p = 0.95$).

$$\text{Response} \sim \text{VOT} * \text{Block} * \text{Condition} + (1 + \text{F0} + \text{Block} + \text{Condition} \mid \text{Listener}) \quad (6)$$

The results are presented in Regression Table 6 (Appendix) and Fig. 8 illustrates predicted probability of [p] responses and standard error of these predictions across blocks and generalization conditions. As in Experiments 1 and 2, robust changes in the influence of F0 across blocks was evident for the Experienced frame (Fig. 8). The critical F0 × Block interaction was significant only from Canonical 1 to Reverse (F0*Canonical1: $\hat{\beta} = 0.39$, SE = 0.20, z = 2.00, p = 0.05; F0*Canonical2: $\hat{\beta} = 0.26$, SE = 0.19, z = 1.39, p = 0.16), indicating that there was an attenuation of F0 influence in the Generalization condition (the reference level) from Canonical 1 to Reverse block, but not from Reverse to Canonical 2 block. This effect interacted with Condition indicating the effect was stronger in the Experienced condition, but the difference between Experienced and Generalization frames was marginally significant only from Canonical 1 to Reverse (F0*Canonical1*Generalization: $\hat{\beta} = 0.50$, SE = 0.26, z = 1.88, p = 0.06; F0*Canonical2*Generalization: $\hat{\beta} = 0.21$, SE = 0.25, z = 0.85, p = 0.40).

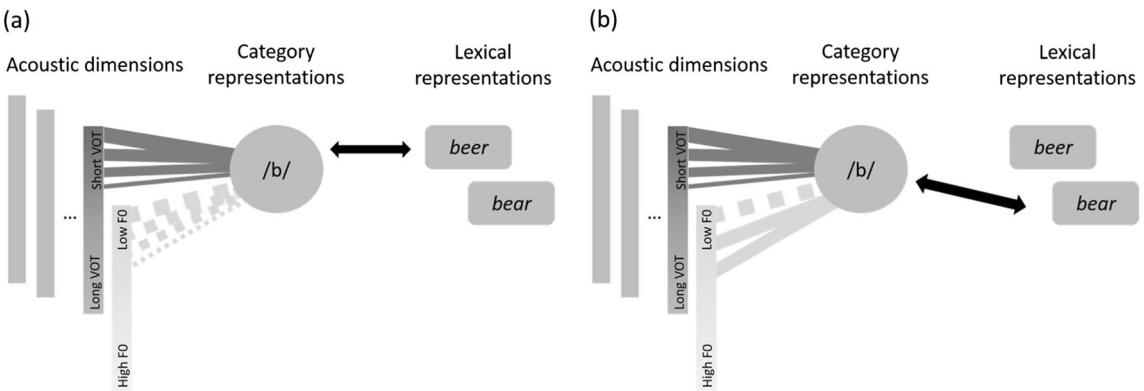

Fig. 9 Schematic illustration of the way acoustic dimensions and phonetic categories may be related for [b]/[p] in *beer* frame and *bear* frame after experience with an accent in *beer* frame. The dashed lines indicates down-weighted activations. F0 dimension is down-weighted robustly for *beer* (**a**). F0 dimension for *bear* (**b**) is down-weighted only through the connection that overlaps with the connection in (**a**)

Similar to Experiments 1 and 2, we found robust perceptual adjustment of F0 influence in the categorization of [b]/[p] when F0/VOT correlation reversed in an artificial accent, but the effect was attenuated in a word frame unexperienced in the accent. Critical difference between the earlier two experiments and the current one was that the artificial accent occurred in multiple word frames conveying that the accent was not limited to one pair of word and was productive. Generalization of the learning in a new frame was still weak. The coefficients for F0 × Block interactions, which estimate the extent of F0 adjustment, in this experiment were 0.54 and 0.14 for F0 × Canonical 1 and F0 × Canonical 2 respectively. The coefficients for the same effects in Experiment 1 and 2 were 0.53 and 0.46 (Experiment 1) and 0.39 and 0.26 (Experiment 2). The comparison of these coefficients do not suggest that F0 down-weighting was more robust in the current experiment compared to the other two. The findings here do not support the prediction that increased variability in word frames impacts generalization of dimension-based statistical learning to a novel word frame.

General discussion

Listeners track acoustic dimensional relationships in online speech processing. When short-term input deviates from the regularities experienced across long-term experience, the contribution of acoustic dimensions to speech categorization is rapidly and dynamically tuned (Idemaru & Holt, 2011; Liu & Holt, 2015). In this *dimension-based statistical learning*, one acoustic dimension, VOT, serves as a teaching signal to convey how the other dimension, F0, maps to voicing categories. Listeners can even simultaneously track dimensional statistics separately across voicing categories of analogous sounds ([b] vs. [p] and [d] vs. [t]) classified typically as belonging to the same phonological class (Idemaru & Holt, 2014). When listeners have evidence that category exemplars should be "binned" separately, for example by experiencing tokens in two different voices or acoustically identical tokens paired with different faces, they can track independent and even opposing short-term input regularities that influence the perceptual weight of input dimensions on speech categorization (Zhang & Holt, 2018). Previous work thus has suggested that dimension-based statistical learning of speech categories is highly specific to the acoustic regularities experienced in the input signal. The current results advance our understanding of the generality and specificity of this learning, underscoring that the system is quite exquisitely sensitive to context and exhibits short-term regularities in speech input in a highly detailed manner.

When listeners experienced accented [b] and [p] with a reversed F0/VOT correlation in a given context, they down-weighted reliance on F0 in categorizing VOT-neutral [b/p] sounds in the context in which they encountered the accent. This rapid learning also generalized to some of the word frames in which they had not experienced the accent. The present experiments confirm this across multiple combinations of experienced words and new words (i.e., *bear/pear – beer/pier* in Experiments 1 and 3*, bear/pear – bill/pill,* and *beer/pier – bill/pill* in Experiment 2). In all*,* we can conclude that dimension-based statistical learning of phonetic categories can transfer across word frames. The results, together with those of previous studies (e.g., Liu & Holt, 2015), indicate that dimension-based statistical learning operates pre-lexically (Idemaru & Holt, 2014). Extant data are consistent with the conclusion that the dynamic adjustments to online speech categorization as a function of short-term statistical regularities experienced across acoustic dimensions are not completely lexically specific. In dimension-based statistical learning the "teaching" signal driving learning arises from the perceptually unambiguous dimension (here, VOT; see Idemaru & Holt, 2011; Liu & Holt, 2015, for discussion). Whereas this contrasts with the lexical teaching signal that drives shifts in speech categorization characteristic of lexically-guided perceptual learning, the pre-lexical locus of learning is consistent across these two perceptual learning phenomena (e.g., Mitterer, Reinisch, & Mcqueen 2018).

The pattern observed in this study furthermore suggests the complex nature of dimension-based statistical learning. Whereas dimension-based statistical learning operates at the pre-lexical level, the extent of learning is modulated by the characteristics of the word frame they appeared.

This observation is consistent with the view of speech representation that allows context-dependent variation in the way acoustic dimensions signal speech categories (Fig. 9). If the way with which F0 and VOT information is communicated to the category [b] varies across different contexts, for example, across *beer* and *bear,* there may be only partial overlap between the two patterns of connection between acoustic dimensions and speech representation. In Fig. 9, two different patterns of connection between the F0 dimension and category representation across *beer* and *bear* are expressed by different numbers of lines connecting the two levels. It follows then that dimension-based statistical learning of speech categories is transferred from one context to another only through the existing overlapping connections between the acoustic dimensions and the speech category. This would result in weaker generalization to a new context. There is also a possibility that diphone, rather than a phone or allophone, is the unit with which dimension-based statistical learning operates. If this is the case, it would explain that generalization from one diphone to another is weaker. Future work is needed to test this possibility.

We considered the possibility that a factor affecting the connections between the acoustic dimensions and the

speech category (thus affecting dimension-based statistical learning) may be the variability of context in which speech categories with an accent are encountered (Experiment 3). More specifically, we predicted that if listeners experience the accent in multiple word frames, it may facilitate more robust generalization. In the model described here, if listeners experience the accent in more word frames, listeners experience more variation in the way in which the acoustic dimensions signal the speech category. One expects, then, dimension-based statistical learning (i.e., adjusting the weight of the connection, as we hypothesized above) in variable contexts would result in more general learning, leading to robust generalization to a novel context. Our results indicated that experiencing the accent in four different word frames spoken by the same talker does not lead to a greater extent of perceptual adjustment than experiencing the accent in a single word frame, indicating that variability in short-term experience, as implemented in the current study, is not effective in driving generalization.

The current findings are not inconsistent with the prior work on generalization of perceptual learning. Studies investigating lexically guided perceptual learning have reported lack of generalization across manner of articulation (Reinisch & Mitterer 2016), allophonic variants (Mitterer, Scharenborg, & McQueen, 2013), and [b]/[d] in different vowel contexts (Reinisch et al., 2014). From the dimension-based statistical learning paradigm, we have reported cases of generalization in vowel categorization. Liu and Holt (2015) tested perceptual adjustment of vowel duration, as a secondary cue, in the categorization of [ɛ]/[æ], and tested its generalization from the *setch-satch* frame to the *set-sat* frame. Unlike the present results, they found robust generalization. An interesting detail of the methodology in Liu and Holt (2015) is that the [sɛ] portion of the *set* and *setch* stimuli, and the [sæ] portion of the *sat* and *satch* stimuli were physically identical, with the *setch-satch* stimuli being created from the *set-sat* as the base. Furthermore, when generalization was tested with a new set of male-sounding *set-sat* generalization stimuli that were acoustically modified from the original female *set-sat* stimuli, generalization was weaker. The current results and those of Liu and Holt (2015) together seem to suggest that the gradedness with which dimension-based statistical learning is applied to a new token of a speech category (e.g., from a token of [p] to another token of [p], and from a token of [ɛ] to another token of [ɛ]) may be influenced by the acoustic similarities across the tokens.

Conclusion

Not only are listeners able to track distributional statistics of acoustic dimensions that define speech categories, they can also track separate statistics across speech categories (e.g., [b]/[p] vs. [d]/[t]) that are considered to belong to the same phonological categories (i.e., stop voicing; Idemaru & Holt, 2014). Furthermore, they seem to be sensitive to the subtly different ways that acoustic dimensions are related to speech categories arising from contextual differences. The present findings suggest the rich and complex nature of speech representation: context-induced variation in the way acoustic dimensions inform speech categories is preserved. Such variation may include the variation in values along the acoustic dimensions, the weight (or strength) of the connection between the acoustic dimensions and associated speech categories, and sensitivity of the connection to adjustment through learning across short-term regularities in the ambient input.

So far, dimension-based statistical learning has shown a remarkable specificity. Learning does not transfer across [b/p] and [d/t] (Idemaru & Holt, 2014). They only transfer the narrowly overlapping properties of the same [b]/[p] sounds across varying contexts, even when the [b] and [p] are produced by the same speaker (as in the case of all experiments presented in this study). Perhaps this makes sense. A speaker may make an idiosyncratic error in producing [p] that does not reoccur any place else. It is certainly not efficient for the auditory system to generalize completely whenever it encounters a deviation from the long-term representation, as such deviation may include one-time, idiosyncratic cases or a pattern specific to a context. However, it also seems inefficient not to generalize completely if there is enough evidence for the deviation being consistent and prevalent. Future research determining the conditions under which the complete generalization occurs will advance understanding of the processes by which speech categories are abstracted across varying contexts. Understanding perceptual adjustment to speech categories as a function of short-term input regularities can inform us about the nature of the long-term representation of speech categories.

Open Practices Statement The data and the analysis code are made available on Open Science Framework.

Author note This research was supported by the National Institutes of Health (R01DC004674), the National Science Foundation (0746067), and Research, Innovation and Graduate Education, University of Oregon. A portion of this work was presented at the 18th International Congress of Phonetic Sciences meeting in August 2015. We thank Christi Gomez and Sara King for running the experiments.

Appendix

Regression Table 1 (Experiment 1: Baseline categorization)

	Est.	S.E.	z val.	p
(Intercept)	-0.72	0.09	-7.94	0.00
VOT	0.28	0.01	20.57	0.00
F0.sum1	0.89	0.06	15.98	0.00
Group.sum1	-0.10	0.09	-1.14	0.25
Pair.sum1	-0.40	0.05	-7.96	0.00
VOT:F0.sum1	-0.03	0.01	-4.03	0.00
VOT:Group.sum1	-0.00	0.01	-0.07	0.94
F0.sum1:Group.sum1	0.11	0.05	2.05	0.04
VOT:Pair.sum1	0.05	0.01	4.99	0.00
F0.sum1:Pair.sum1	-0.22	0.05	-4.31	0.00
Group.sum1:Pair.sum1	-0.04	0.05	-0.87	0.38
VOT:F0.sum1:Group.sum1	-0.02	0.01	-2.73	0.01
VOT:F0.sum1:Pair.sum1	0.00	0.01	0.44	0.66
VOT:Group.sum1:Pair.sum1	-0.00	0.01	-0.23	0.82
F0.sum1:Group.sum1:Pair.sum1	0.01	0.05	0.13	0.89
VOT:F0.sum1:Group.sum1:Pair.sum1	-0.01	0.01	-1.89	0.06

Note: The reference levels are F0 (Low), Group (Bear), and Pair (*beer-pier*)

Regression Table 2 (Experiment 1: Categorization of test stimuli during exposure to accent)

	Est.	S.E.	z val.	p
(Intercept)	-1.65	0.15	-10.67	0.00
F0.sum1	0.75	0.10	7.76	0.00
BlockCanoncl1	0.04	0.12	0.29	0.77
BlockCanoncl2	0.22	0.12	1.88	0.06
CondExprncd	0.73	0.16	4.51	0.00
Group.sum1	-0.18	0.15	-1.14	0.25
F0.sum1:BlockCanoncl1	0.47	0.12	3.79	0.00
F0.sum1:BlockCanoncl2	0.15	0.12	1.26	0.21
F0.sum1:CondExprncd	-0.77	0.12	-6.40	0.00
BlockCanoncl1:CondExprncd	-0.34	0.17	-1.98	0.05
BlockCanoncl2:CondExprncd	-0.33	0.16	-2.04	0.04
F0.sum1:Group.sum1	-0.28	0.10	-2.91	0.00
BlockCanoncl1:Group.sum1	-0.23	0.12	-1.90	0.06
BlockCanoncl2:Group.sum1	0.01	0.12	0.10	0.92
CondExprncd:Group.sum1	0.68	0.16	4.26	0.00
F0.sum1:BlockCanoncl1:CondExprncd	1.35	0.17	7.92	0.00
F0.sum1:BlockCanoncl2:CondExprncd	1.51	0.16	9.39	0.00
F0.sum1:BlockCanoncl1:Group.sum1	-0.05	0.12	-0.40	0.69
F0.sum1:BlockCanoncl2:Group.sum1	-0.17	0.12	-1.43	0.15
F0.sum1:CondExprncd:Group.sum1	0.56	0.12	4.65	0.00
BlockCanoncl1:CondExprncd:Group.sum1	0.27	0.17	1.58	0.11
BlockCanoncl2:CondExprncd:Group.sum1	-0.34	0.16	-2.11	0.03
F0.sum1:BlockCanoncl1:CondExprncd:Group.sum1	0.18	0.17	1.06	0.29
F0.sum1:BlockCanoncl2:CondExprncd:Group.sum1	0.31	0.16	1.96	0.05

Note: Canoncl = Canonical, Cond = Condition, Exprncd = Experienced

The reference levels are F0 (Low), Block (Reverse), Condition (Generalization), Group (Bear)

Regression Table 3 (Experiment 2: Baseline categorization)

	Est.	S.E.	z val.	p
(Intercept)	-0.03	0.09	-0.37	0.71
VOT	0.28	0.02	16.28	0.00
F0.sum1	1.04	0.06	16.27	0.00
Group.sum1	0.08	0.08	0.98	0.33
Pair.sum1	-0.90	0.10	-8.95	0.00
Pair.sum2	-0.01	0.10	-0.09	0.93
VOT:F0.sum1	-0.02	0.01	-3.26	0.00
VOT:Group.sum1	0.02	0.01	1.12	0.26
F0.sum1:Group.sum1	0.02	0.06	0.37	0.71
VOT:Pair.sum1	0.05	0.02	2.54	0.01
VOT:Pair.sum2	-0.03	0.02	-1.41	0.16
F0.sum1:Pair.sum1	-0.21	0.07	-2.86	0.00
F0.sum1:Pair.sum2	0.27	0.07	3.78	0.00
VOT:F0.sum1:Group.sum1	-0.01	0.01	-1.25	0.21
VOT:F0.sum1:Pair.sum1	-0.00	0.01	-0.40	0.69
VOT:F0.sum1:Pair.sum2	-0.00	0.01	-0.18	0.85

Note: The reference levels are F0 (Low), Group (Bear), and Pair (*bill-pill*). For Pair, Contrast 1 compared *bill-pill* and *bear-pear*

Regression Table 4 (Experiment 2: Categorization of test stimuli during exposure to accent)

	Est.	S.E.	z val.	p
(Intercept)	1.03	0.17	5.92	0.00
F0.sum1	1.11	0.10	10.84	0.00
BlockCanoncl1	-0.04	0.12	-0.32	0.75
BlockCanoncl2	-0.24	0.12	-2.05	0.04
status.tExprncd	-1.52	0.21	-7.28	0.00
Group.sum1	0.31	0.17	1.80	0.07
F0.sum1:BlockCanoncl1	0.58	0.12	4.94	0.00
F0.sum1:BlockCanoncl2	0.52	0.11	4.54	0.00
F0.sum1:status.tExprncd	-0.88	0.10	-8.48	0.00
BlockCanoncl1:CondExprncd	-0.35	0.16	-2.23	0.03
BlockCanoncl2:CondExprncd	-0.11	0.16	-0.70	0.49
F0.sum1:Group.sum1	-0.08	0.10	-0.74	0.46
BlockCanoncl1:Group.sum1	-0.28	0.12	-2.32	0.02
BlockCanoncl2:Group.sum1	-0.04	0.12	-0.34	0.74
CondExprncd:Group.sum1	0.09	0.21	0.45	0.65
F0.sum1:BlockCanoncl1:CondExprncd	0.86	0.16	5.43	0.00
F0.sum1:BlockCanoncl2:CondExprncd	1.17	0.16	7.32	0.00
F0.sum1:BlockCanoncl1:Group.sum1	0.02	0.12	0.16	0.88
F0.sum1:BlockCanoncl2:Group.sum1	-0.05	0.11	-0.44	0.66
F0.sum1:CondExprncd:Group.sum1	0.22	0.10	2.10	0.04
BlockCanoncl1:CondExprncd:Group.sum1	0.10	0.16	0.62	0.54
BlockCanoncl2:CondExprncd:Group.sum1	-0.04	0.16	-0.25	0.81
F0.sum1:BlockCanoncl1:CondExprncd:Group.sum1	0.19	0.16	1.21	0.23
F0.sum1:BlockCanoncl2:CondExprncd:Group.sum1	0.21	0.16	1.32	0.19

Note: Canoncl = Canonical, Cond = Condition, Exprncd = Experienced

The reference levels are F0 (Low), Block (Reverse), Condition (Generalization), Group (Bear)

Regression Table 5 (Experiment 3: Baseline categorization)

	Est.	S.E.	z val.	p
(Intercept)	-0.72	0.15	-4.93	0.00
VOT	0.28	0.02	11.60	0.00
F0.sum1	0.80	0.06	12.82	0.00
Pair.sum1	-0.29	0.03	-9.36	0.00
VOT:F0.sum1	-0.03	0.01	-3.60	0.00
VOT:Pair.sum1	0.03	0.00	7.39	0.00
F0.sum1:Pair.sum1	-0.27	0.03	-8.91	0.00
VOT:F0.sum1:Pair.sum1	0.00	0.00	0.90	0.37

Note: The reference levels are F0 (Low), and Pair (*beer-pier*)

Regression Table 6 (Experiment 3: Categorization of test stimuli during exposure to accent)

	Est.	S.E.	z val.	p
(Intercept)	-2.78	0.38	-7.35	0.00
F0.sum1	0.47	0.19	2.47	0.01
BlockCanoncl1	-0.11	0.25	-0.44	0.66
BlockCanoncl2	-0.02	0.25	-0.09	0.93
CondExprncd	0.96	0.25	3.79	0.00
F0.sum1:BlockCanoncl1	0.39	0.20	2.00	0.05
F0.sum1:BlockCanoncl2	0.26	0.19	1.39	0.16
F0.sum1:CondExprncd	0.44	0.18	2.48	0.01
BlockCanoncl1:CondExprncd	0.04	0.27	0.15	0.88
BlockCanoncl2:CondExprncd	0.03	0.26	0.13	0.90
F0.sum1:BlockCanoncl1:CondExprncd	0.50	0.26	1.88	0.06
f0.sum1:BlockCanonical2:Cond.texperienced	0.21	0.25	0.85	0.40

Note: Canoncl = Canonical, Cond = Condition, Exprncd = Experienced

The reference levels are F0 (Low), Block (Reverse), Condition (Generalization)

References

Abramson, A. S., & Lisker, L. (1985). Relative power of cues: F0 shift versus voice timing. Phonetic linguistics: Essays in honor of Peter Ladefoged, 25–33.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. arXiv preprint arXiv:1406.5823.

Bertelson, P., Vroomen, J., & de Gelder, B. (2003). Visual Recalibration of Auditory Speech Identification A McGurk Aftereffect. Psychological Science, 14(6), 592-597.

Boersma, P. & Weenink, D. (2017). Praat: doing phonetics by computer [Computer program]. Version 5.0, retrieved from http://www.praat.org/

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

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.

Cutler, A., McQueen, J. M., Butterfield, S., & Norris, D. (2008). Prelexically-driven perceptual retuning of phoneme boundaries. In J. Fletcher, D. Loakes, M. Wagner, & R. Goecke. (eds.), Proceedings of Interspeech 2008 (2056). Brisbane, Australia: ISCA.

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

Haggard, M., Ambler, S., & Callow, M. (1970). Pitch as a voicing cue. Journal of the Acoustical Society of America, 47, 613-617.

Hazan, V., & Barrett, S. (2000). The development of phonemic categorization in children aged 6-12. Journal of Phonetics, 28(4), 377–396.

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.

Idemaru, K., Holt, L. L., Seltman, H. (2012). Individual differences in cue weights are stable across time: the case of Japanese stops lengths. The Journal of the Acoustical Society of America, 132(6), 3950-3964.

Idemaru, K., & Holt, L. L. (2013). The long developmental trajectory of children's perception and production of English /r/-/l/. Journal of the Acoustical Society of America, 133, 4232 – 4246. doi: https://doi.org/10.1121/1.4802905

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

Idemaru, K. & Guion, S. G. (2008). Acoustic covariants of length contrast in Japanese stops. Journal of International Phonetic Association, 38(2), 167-186. doi: https://doi.org/10.1017/S0025100308003459

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. Journal of memory and language, 59(4), 434-446.

Kong, E. J., & Edwards, J. (2016). Individual differences in categorical perception of speech: Cue weighting and executive function. Journal of Phonetics, 59, 40-57.

Kraljic, T., & Samuel, A. G. (2005). Perceptual learning for speech: Is there a return to normal?. Cognitive psychology, 51(2), 141-178.

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

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

Lehet, M., & Holt, L. L. (2017). Dimension-based statistical learning affects both speech perception and production. Cognitive science, 41, 885-912.

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

Lotto, A.J., Sato, M., & Diehl, R.L. (2004). Mapping the task for the second language learner: The case of Japanese acquisition of /r/ and /l/. From Sound to Sense: 50+ Years of Discoveries in Speech Communication, edited by J. Slifka, S. Manuel, & M. Matthies. Electronic conference proceedings, 181-186.

Maye, J., & Gerken, L. (2001). Learning phonemes: How far can the input take us. In Proceedings of the 25th annual Boston University conference on language development (Vol. 1, p. 480). Somerville, MA: Cascadilla Press.

McQueen, J. M., Cutler, A., & Norris, D. (2006). Phonological abstraction in the mental lexicon. *Cognitive Science, 30*(6), 1113–1126.

Mitterer, H., Cho, T., & Kim, S. (2016). What are the letters of speech? Testing the role of phonological specification and phonetic similarity in perceptual learning. Journal of Phonetics, 56, 110-123.

Mitterer, H., Reinisch, E., & McQueen, J. M. (2018). Allophones, not phonemes in spoken-word recognition. Journal of Memory and Language, 98, 77-92.

Mitterer, H., Scharenborg, O., & McQueen, J. M. (2013). Phonological abstraction without phonemes in speech perception. Cognition, 129(2), 356-361.

McMurray, B., & Aslin, R. N. (2005). Infants are sensitive to within-category variation in speech perception. Cognition, 95(2), B15-B26.

Nusbaum, H. C., & Morin, T. M. (1992). Paying attention to differences among talkers. Speech perception, production and linguistic structure, 113-134.

Nittrouer, S. (1992). Age-related differences in perceptual effect of formant transitions within syllables and across syllable boundaries. Journal of Phonetics, 20, 1–32.

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. Cognitive Psychology, 47(2), 204-238.

R Core Team. (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://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(2), 539.

Reinisch, E., & Mitterer, H. (2016). Exposure modality, input variability and the categories of perceptual recalibration. Journal of Phonetics, 55, 96-108.

Reinisch, E., Wozny, D. R., Mitterer, H., & Holt, L. L. (2014). Phonetic category recalibration: What are the categories?. Journal of phonetics, 45, 91-105.

Schertz, J., T. Cho, A. Lotto, and N. Warner. (2016). Individual differences in perceptual adaptability of foreign sound categories. *Attention, Perception, & Psychophysics, 78*(1), 355-367.

Theodore, R. M., Myers, E. B., & Lomibao, J. A. (2015). Talker-specific influences on phonetic category structure. *The Journal of the Acoustical Society of America, 138*(2), 1068–1078.

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.

Whalen, D. H., Abramson, A. S., Lisker, L., & Mody, M. (1993). F0 gives voicing information even with unambiguous voice onset times. The Journal of the Acoustical Society of America, 93, 2152-2159.

Zhang, X. & Holt, L. L. (2018). Simultaneous tracking of co-evolving distributional regularities in speech. *Journal of Experimental Psychology: Human Perception & Performance, 44*(11), 1760.