Atten Percept Psychophys
DOI 10.3758/s13414-015-0987-1

Individual differences in perceptual adaptability of foreign sound categories

Jessamyn Schertz[1] · **Taehong Cho**[2] · **Andrew Lotto**[3] · **Natasha Warner**[4]

Abstract Listeners possess a remarkable ability to adapt to acoustic variability in the realization of speech sound categories (e.g., different accents). The current work tests whether non-native listeners adapt their use of acoustic cues in phonetic categorization when they are confronted with changes in the distribution of cues in the input, as native listeners do, and examines to what extent these adaptation patterns are influenced by individual cue-weighting strategies. In line with previous work, native English listeners, who use voice onset time (VOT) as a primary cue to the stop voicing contrast (e.g., 'pa' vs. 'ba'), adjusted their use of f0 (a secondary cue to the contrast) when confronted with a noncanonical "accent" in which the two cues gave conflicting information about category membership. Native Korean listeners' adaptation strategies, while variable, were predictable based on their initial cue weighting strategies. In particular, listeners who used f0 as the primary cue to category membership adjusted their use of VOT (their secondary cue) in response to the noncanonical accent, mirroring the native pattern of "downweighting" a secondary cue. Results suggest that non-native listeners show native-like sensitivity to distributional information in the input and use this information to adjust categorization, just as native listeners do, with the specific trajectory of category adaptation governed by initial cue-weighting strategies.

Keywords Categorization · L2 speech perception · Perceptual learning · Psycholinguistics

✉ Jessamyn Schertz
jessamyn.schertz@utoronto.ca

[1] Centre for French and Linguistics, University of Toronto, 1265 Military Trail, HW314, Toronto, ON M1C 1A4, Canada

[2] Hanyang Phonetics and Psycholinguistics Laboratory, Department of English Language and Literature, College of Humanities, Hanyang University, Seoul 133-791, Korea

[3] Department of Speech, Language and Hearing Sciences, The University of Arizona, PO Box 210071, Tucson, AZ 85721, USA

[4] Department of Linguistics, The University of Arizona, PO Box 210025, Tucson, AZ 85721, USA

The massive amount of variability inherent to speech requires that listeners make rapid, dynamic adjustments to their definitions of sound categories. Listeners are regularly confronted with dialects and accents in which the "same" sounds are realized differently, and even talkers with similar accents produce the same sounds with different acoustic realizations, due to anatomical differences in the vocal tract. While the details of how listeners resolve the "lack of invariance" problem remain elusive, what is clear is that listeners possess a remarkable amount of perceptual flexibility, rapidly accommodating to foreign accents (e.g., Clarke & Garrett, 2004; Bradlow & Bent, 2008; Baese-Berk, Bradlow, & Wright, 2013), dialectal variation (e.g., Sumner & Samuel, 2009; Trude & Brown-Schmidt, 2012), and degraded speech (e.g., Davis, Johnsrude, Hervais-Adelman, Taylor, & McGettigan, 2005). In contrast to the plasticity found in studies examining native listeners, work on non-native (L2) speech perception has, for the most part, focused on the notoriously stubborn native-language (L1) constraints on L2 sound category formation and perception (see Flege, 1995; Best, 1995, among others). This apparent asymmetry between L1 and L2 perceptual flexibility suggests the possibility that L2 listeners may employ qualitatively or quantitatively different adaptation strategies than L1 listeners. In the current work, we examine this possibility by comparing how native and non-native listeners adapt their phonetic categorization strategies in response to different

Published online: 24 September 2015

"accents" that vary in the distribution of acoustic cues defining phonetic categories.

Speech perception can be thought of as an example of a general auditory categorization task, with sound categories being mapped onto a multi-dimensional acoustic space (Goudbeek, Smits, Cutler, & Swingley, 2005; Goudbeek, Swingley, & Smits, 2009; Holt & Lotto, 2008, 2010). Speech sounds contrast on many acoustic dimensions (or "cues"), and listeners give different "weights" to these dimensions. For example, the primary cue to the English stop voicing contrast (/p/ vs /b/, /t/ vs /d/, and /g/ vs /k/) is voice onset time (VOT), or the time lapse between the release of the stop closure and the onset of voicing in the following vowel (e.g., Lisker & Abramson, 1964); however, other secondary cues, including fundamental frequency (f0) at vowel onset, also define the contrast, albeit less reliably (e.g., House & Fairbanks, 1953; Kingston & Diehl, 1994; Francis, Kaganovich, & Driscoll-Huber, 2008; Kingston, Diehl, Kirk, & Castleman, 2008; Llanos, Dmitrieva, Shultz, & Francis, 2013). Native English speakers' productions of voiceless stops /p, t, k/ have longer VOTs than voiced stops /b, d, g/; on average, voiceless stops are also produced with slightly higher f0 at vowel onset than voiced stops. However, in contrast to VOT, which consistently separates productions of voiced and voiceless stops, there is a large amount of overlap in the use of f0 between the two categories. These distributional patterns present in speakers' productions are reflected in listeners' perception of the contrast, which they distinguish primarily using VOT; secondary cues like f0 can influence categorization decisions, but do so to a much lesser extent (e.g., Whalen, Abramson, Lisker, & Mody, 1993; Francis et al., 2008).

Listeners adapt their use of these phonetic dimensions in response to many factors, including the acoustic properties of surrounding auditory stimuli (e.g., selective adaptation: Eimas & Corbit, 1973; contrast effects: Diehl, Elman, & McCusker, 1978) and the distributional properties of the relevant dimensions (e.g., Clayards, Tanenhaus, Aslin, & Jacobs, 2008, cf. Kleinschmidt & Jaeger, 2015). Whereas listeners may derive stable, "prototypical" cue weights for a contrast based on their long-term experience with acoustic realizations of these sounds aggregated across many speakers and tokens, any given set of weights is unlikely to be optimal for one particular speaker. A successful listener must therefore be able to adapt rapidly to the unpredictable idiosyncrasies of a new speaker. Several lines of research have investigated specific ways in which listeners can be induced to shift their sound categories based on higher-level semantic or lexical information (see Samuel & Kraljic, 2009 for a review). For example, Norris, McQueen, and Cutler (2003) showed that after words containing an ambiguous sound between [f] and [s] that were lexically disambiguated to be [f], listeners were more likely to subsequently characterize ambiguous sounds on an [f]–[s] continuum as [f] (see also Kraljic & Samuel, 2005; Eisner & McQueen, 2005; Cutler, McQueen, Butterfield, & Norris, 2008, among many others). Along the same lines, in work by Maye, Aslin, and Tanenhaus (2008), after listening for 20 min to a synthesized English talker whose front vowels were shifted categorically lower in the vowel space, listeners adapted their vowel categories to this idiosyncratic accent. The sort of contextual information that can be used by listeners goes beyond the acoustic level: Bertelson, Vroomen, and de Gelder (2003) demonstrated that exposure to audiovisual information during speech can influence subsequent auditory categorization. All of this work demonstrates that listeners shift their criterion for a category boundary on a given dimension after implicitly "learning" idiosyncratic use of cues from contextual information.

In contrast to the "object-based" learning discussed above, in which lexical (or other higher-level) information disambiguates non-prototypical phonetic characteristics, Idemaru and Holt (2011) demonstrated "dimension-based" statistical learning, in which listeners modify their use of a secondary acoustic dimension defining a given sound contrast based on its relationship with a primary acoustic dimension in the input. In particular, Idemaru and Holt (2011) investigated whether listeners would adjust their use of f0 when categorizing the English stop voicing contrast based on short-term changes in the correlation of VOT (the primary cue) and f0 (the secondary cue) in the input. Distributions of stimuli with unambiguously long or short VOT (/p/ vs /b/) and varying values of f0 were presented to listeners in different blocks (see Fig. 2 below for a schematic). One block was characterized by the canonical English correlation of f0 and VOT: tokens with long VOT (i.e., /p/) had high f0, while tokens with short VOT (i.e., /b/) had low f0. This "Canonical" block was followed by a "Reversed" block showing the opposite correlation (long VOT with *low* f0 and short VOT with *high* f0). There were no clues or instructions to denote the introduction of this artificial "accent" and since all other aspects of the talker remained the same, participants did not consciously note the change. Despite this lack of explicit knowledge, listeners modified their use of f0 across the two blocks, as demonstrated by their response patterns when categorizing ambiguous /p/ ~/b/ stimuli. In the context of the Canonical accent, listeners made use of f0 to categorize stimuli with intermediate values of VOT (high f0 elicited more /p/ responses). In the Reversed block, on the other hand, f0 had no effect on categorization responses, suggesting that listeners "downweight" their reliance on a secondary cue (f0) when confronted with noncanonical use of the cue in the input. The authors concluded that listeners recruit what they know to be a more reliable dimension (VOT) as the basis for learning about the distribution of a less reliable dimension (f0) within a given accent, then adjust their use of the secondary dimension accordingly.

A general formulation of dimension-based learning requires that listeners be (1) *sensitive* to the statistical distribution of secondary phonetic dimensions in the input (even when attention to a given dimension is not directly necessary for the task) and (2) *adaptable* with respect to these dimensions. Idemaru and Holt's (2011) native-listener participants demonstrated both of these characteristics: they paid attention to how f0 was used, even when VOT gave unambiguous information about category membership, and modulated their use of f0 when it did not match with the canonical distribution. Whether or not non-native speech perception is characterized by comparable sensitivity and adaptability is an open question. The fact that L1 phonetic patterns exert a strong and (to some extent) predictable influence on L2 perception is uncontroversial. On the other hand, several studies have demonstrated that L2 listeners are able to adjust their cue-weighting strategies in the context of training paradigms designed to direct attention toward relevant acoustic dimensions by exaggerating the contrast (Iverson, Hazan, & Bannister, 2005; Kondaurova & Francis, 2010; Escudero, Benders, & Wanrooij, 2011) or to direct attention *away* from less relevant dimensions via increased variability (Iverson et al., 2005; Kondaurova & Francis, 2010; Lim & Holt, 2011; cf. Holt & Lotto, 2006). However, these tasks are for the most part characterized by explicit feedback over extended training (though the feedback in Lim and Holt (2011) was indirect) Furthermore, in the absence of direct control groups, it is not clear whether or how L1 listeners would also shift categorization in these sorts of paradigms.

In general, the types of adaptation addressed in L1 vs. L2 perceptual learning studies are conceptualized in qualitatively different ways (e.g. L1 category "tuning" vs. L2 "training"). The extent to which these actually constitute distinct processes is an empirical question, and one which is complicated by the fact that the modifications required to shift from non-native to native-like cue weighting strategies are usually much more extensive than the fine-grained perceptual "tuning" elicited by L1 adaptation studies. For example, the well-known difficulty distinguishing the English /r/-/l/ contrast for native Japanese listeners is attributed to the fact that they do not use F3 as a cue to the contrast, as native English listeners do (e.g., Miyawaki et al., 1975; Yamada & Tohkura, 1990; Iverson et al., 2003). L2 perceptual development thus depends on shifting attention to an entirely new dimension (e.g., Francis & Nusbaum, 2002), while L1 category adaptation work has generally focused on small criterion shifts on an already-used dimension (e.g., Norris et al., 2003). The discrepancy between the types of adaptation usually examined in L1 vs L2 perceptual learning therefore makes it difficult to determine whether any differences in native vs non-native perceptual learning and plasticity found in previous work reflect fundamentally different processes of accommodation, or whether instead they fall out from the different types of adaptation generally targeted for the two groups in laboratory tasks. Recent results from Reinisch, Weber and Mitterer (2013) and Schuhmann (2014) lend support to the latter hypothesis: L2 listeners showed similar shifts on /f/-/s/ continua as L1 listeners, suggesting that L1 and L2 listeners use similar processes for phonetic category adjustment. The languages used in these studies (Dutch and German in Reinisch et al., 2013; German and English in Schuhmann, 2014) have very similar phonetic realizations of the target contrast (/f/-/s/), leaving open the question of whether similar retuning occurs when the L2 contrast does not have a close phonetic match in the L1. In the current work, the target stop contrast is realized very differently in the listeners' L2 (English) than it is in their L1 (Korean), allowing us to test the generality of adaptation processes in native and non-native perception.

In addition to comparing adaptation of cue weights in L1 vs. L2, the examination of the L2 learners' adaptation provides an opportunity to test one of the main hypotheses of the account proposed by Idemaru and Holt (2011). In their description of dimension-based learning, Idemaru and Holt (2011) suggest that the primary cue to English stop voicing (VOT) serves as a learning signal for weighting of the secondary cue (f0). This could occur through direct comparison of the two dimensions or by an error signal coming from the category representation activated by the primary cue (Guediche, Blumstein, Fies, & Holt, 2014). A prediction of this primary-cue-based learning account is that the pattern of learning (adaptation) should depend on the initial relative cue weights of the listener. This prediction is hard to test with L1 listeners because there is little inter-individual variability in the primacy of VOT as a cue to the English voicing contrast. However, the well-attested variability in non-native (L2) sound perception provides a good potential test ground for the hypothesis.

This hypothesis is further motivated by several previous findings of differential adaptation patterns based on differences in initial categorization strategies. Chandrasekaran, Sampath, and Wong (2010) showed that native English listeners' success in a perceptual learning task targeting Mandarin tonal contrast was correlated with initial attention to cues: in particular, listeners who paid more attention to the trajectory of f0 during a pre-test showed larger effects of training (while training-related effects were not related to listeners' initial attention to f0 height). Similarly, in work by Wanrooij, Escudero, and Raijmakers (2013), L2 Dutch learners responded differently to distributional training on the / /-/a:/ contrast based on whether or not they used spectral (i.e., F1 and F2) cues to the contrast prior to training. Turning to native listeners, Sawusch and Nusbaum (1983) showed that the same pair of stimuli elicited different directions of contrast effects from different listeners, and that the direction of the effect was predictable from listeners' initial categorization of the sounds.

In previous work (Schertz, Cho, Lotto, & Warner, 2015), we found substantial variability in native Korean listeners' cue weighting strategies in distinguishing their L2 English stop contrast: while some Korean listeners used primarily VOT to distinguish the contrast in a forced-choice task, like native English listeners do, most either used primarily f0, or made use of both dimensions (requiring both long VOT and high f0 to categorize stimuli as voiceless /p/). This tendency to rely on f0 likely stems from the fact that the three-way stop contrast in Korean relies heavily on both VOT and f0 (e.g. Cho, Jun, & Ladefoged, 2002; Lee & Jongman, 2012). Interestingly, Korean speakers vary both VOT and f0 to an equal extent when distinguishing their L2 English stop contrast in *production* (Schertz et al., 2015), but this cue use is not necessarily reflected in their *perception*. The factors underlying these differences in phonetic structure are not yet known; however, recent work by Kong and Yoon (2013) suggests that listeners' level of English proficiency plays a role, with higher-proficiency speakers using f0 less (i.e. in a more native-like way) than lower-proficiency speakers. Another potential source of variability is the multiple options for mapping the English contrast onto the three-way Korean contrast (e.g. Park & de Jong, 2008). Regardless of the sources of these differences, the different cue weighting strategies (i.e., different L2 listeners consider different dimensions as primary) allow us to test the hypothesis that phonetic category modification can occur as a function of one of the dimensions acting as an anchor, and that this anchor dimension is based on listener-specific internal organization of acoustic cues to category membership.

The current work aimed to address two issues. First, we examined whether non-native listeners show native-like category adaptation strategies when confronted with changes in the distributional properties of acoustic dimensions via a direct comparison between L1 English and L2 English/L1 Korean listeners. Second, we tested the hypothesis that listeners adjust their use of secondary cues to category membership by using a reliable dimension as an "anchor" to extract information about other, less reliable, dimensions. The individual variability that often underlies L2 perception, and in particular the expectation, based on previous work, that L2 Korean listeners will show different cue-weighting strategies for the English stop voicing contrast, allows for a robust test of the prediction that individual differences in categorization strategies lead to categorically different adaptation patterns.

To test these questions, we exposed L1 Korean/L2 English listeners and a control group of L1 English listeners to English sentences containing target syllables beginning with word-initial stops manipulated to covary on two dimensions, VOT and f0, following a modified paradigm of Idemaru and Holt (2011). Although the range and distribution of stimuli along each of the two dimensions remained constant throughout the experiment, the relationship between the dimensions varied by block. In "Canonical" blocks, consistent with the canonical English voicing contrast, VOT and f0 covaried in a positive direction (e.g., long VOT was paired with *high* f0), while in "Reversed" blocks, they covaried in the opposite direction (e.g. long VOT was paired with *low* f0).

Following Idemaru and Holt (2011), we expected native English listeners to use VOT as the dominant anchor dimension, adapting their use of f0 (their secondary dimension) in categorizing stimuli with intermediate values of VOT (which should be ambiguous with respect to category membership). Based on Korean perception data reported in Schertz et al. (2015), we expected individual Korean listeners to use different strategies for distinguishing the contrast, with some relying primarily on VOT, some relying primarily on f0, and some relying on the two dimensions to a similar extent. If the same processes drive adaptation in non-native sound categories, regardless of which dimension is dominant, then we would expect different (but symmetrical) patterns of adaptation for the Korean listeners, with the specific pattern determined by these initial individual categorization patterns. To test the hypothesis that listeners use a dominant dimension as an anchor or learning signal, our main comparison of interest is between native English listeners (who use primarily VOT) and those Korean listeners who use primarily f0 (with VOT as a secondary cue). For these Korean listeners, we expect to see the mirror image of native English listeners' behavior, using *f0* as their anchor dimension and adapting their use of *VOT* when categorizing stimuli with intermediate values of f0.

On the other hand, non-native listeners may not employ native-like category adaptation strategies. Although listeners have been shown to be sensitive to distributional information in their L2, these findings have come primarily from category training tasks with explicit feedback (though see Lim & Holt, 2011 for improved L2 categorization on a videogame task without direct feedback), and these tasks differ substantially from those examining L1 category tuning. Although recent work suggests that L2 listeners show lexically guided phonetic tuning (Reinisch et al., 2013), this has been shown only for phonetic categories that are virtually identical across the two languages. Furthermore, although we expect most Korean listeners to rely on f0 more than VOT for the English contrast, VOT may still play a significant role, given that VOT is the most reliable indicator of English stop category membership in Koreans' *productions* of their L2 English contrast (Schertz et al., 2015).

Methods

Participants

Forty native Korean-speaking undergraduate students at Hanyang University in Seoul (20 male, 20 female, ranging in age from 19 to 29 years) were paid for their participation. All

Korean participants had learned English in school (beginning at a mean age of 9.6 years), but none used it on a regular basis. A control group of 23 native English listeners from the University of Arizona (10 male, 13 female, ranging in age from 18 to 26 years) received course credit for their participation. All listeners reported normal hearing with no speech disorders.

Stimuli

Stimuli consisted of stop-initial target syllables (i.e., a [pa]-[ba] series) embedded in an English carrier sentence. The series of target syllables was created by manipulating a female native English speaker's production of the syllable [pa]. Using the acoustic analysis software Praat (Boersma & Weenink, 2011), a series of stop-initial syllables varying in VOT and f0 at vowel onset was created, resulting in a set of stimuli spanning a two-dimensional acoustic space: nine steps of VOT, ranging from –20 to 50 ms by nine steps of f0, ranging from 160 to 240 Hz, for a total of 81 stimuli (the ranges for each dimension were chosen based on native Korean listeners' categorization crossover points from previous work, Schertz et al., 2015). Waveforms and spectrograms of stimuli at the two endpoints of the VOT series, as well as a schematic of the f0 contours in the stimulus range, are given in Fig. 1. To create stimuli with positive VOT values, aspiration duration was manipulated using the time-domain f0-synchronous-overlap-and-add algorithm (TD-PSOLA, Moulines & Charpentier, 1990) as implemented in Praat. This algorithm manipulates duration of a sound by remapping portions of the original signal onto a new signal, repeating windowed portions of the signal at regular intervals to increase duration and removing portions to decrease duration (for voiced sounds, the windows are based on f0 periods, whereas for voiceless sounds, portions of the sound are simply copied in order to increase duration). To create tokens with negative VOT (i.e., prevoicing), aspiration duration was set to zero (as described above), then consecutive periods of prevoicing were added before the stop burst. F0 was also manipulated using the TD-PSOLA algorithm (as implemented in Praat) to remain at the desired value for the first half of the vowel, then fell linearly to 140 Hz for all stimuli (see Fig. 1). Each syllable was then embedded in an English carrier phrase ("I say [*target syllable*]"), recorded by the same speaker (the carrier phrase was included to keep listeners in an "English mode").

As in Idemaru and Holt (2011), these stimuli were distributed among three blocks (Neutral, Canonical, and Reversed); the distribution of stimuli is shown in Fig. 2. The Neutral block consisted of "baseline" stimuli spanning the entire covarying VOT-f0 stimulus space (81 stimuli, repeated twice for a total of 162 trials in the block). This block was used to orient listeners to the acoustic space and was not included in any further analyses. The two types of test blocks (Canonical and Reversed) contained a subset of these baseline stimuli. In each block, ten "covarying" stimuli ("exposure stimuli" in Idemaru & Holt, 2011) had extreme values of VOT and f0. Each block also contained two "ambiguous-VOT" stimuli (analogous to the "test stimuli" in Idemaru & Holt, 2011) with intermediate values of VOT (15 ms) with relatively low or high f0 (170 or 230 Hz), and two "ambiguous-f0" stimuli (not included in Idemaru & Holt, 2011) with intermediate values of f0 (200 Hz) and relatively high and low values for VOT (–11 and 41 ms). The four "ambiguous" stimuli were identical in the Canonical and Reversed blocks; the two types of blocks differed only in the correlation of f0 and VOT in the covarying stimuli. In the Canonical block, the covarying stimuli were modeled after the canonical English pattern, such that stimuli with *long* VOT (i.e., voiceless stops, 33 ms, 41 ms, or 51 ms) had *high* f0 (220 Hz, 230 Hz, or 240 Hz), while stimuli with *short* VOT (i.e. voiced stops, 2 ms, –11 ms, or –20 ms) had *low* f0 (160 Hz, 170 Hz, or 180 Hz). In the Reversed block, the relationship between VOT and f0 was switched: the Reversed covarying stimuli with long VOT (i.e., voiceless stops) had *low* f0, while those with short VOT (i.e., voiced stops) had *high* f0. In total, each test block contained 140 stimuli consisting of ten randomized repetitions of the 14 stimuli (ten covarying plus four ambiguous); the covarying vs ambiguous stimuli were intermixed within the block and undifferentiated to the listeners. The mixture of the covarying

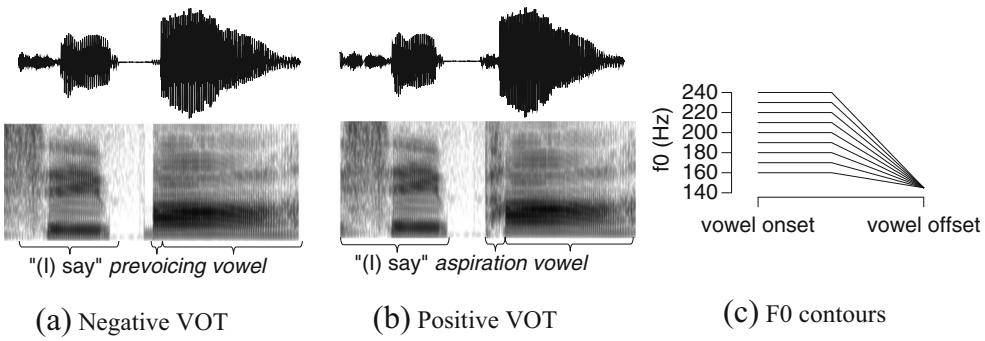


Fig. 1 Waveforms (above) and spectrograms (below) of a stimulus at the endpoints of the voice onset time (VOT) series, with VOT of −20 ms (**a**) and VOT of 50 ms (**b**). Each figure shows the end of the carrier phrase ("say") along with the target syllable. Each of the nine steps of VOT was crossed with the nine f0 contours schematized in (**c**) to create 81 stimuli

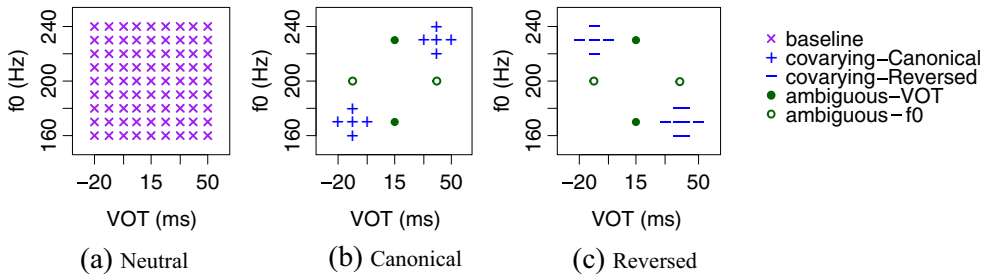


Fig. 2 Distribution of stimuli in the **a** Neutral, **b** Canonical, and **c** Reversed blocks. The stimuli are differentiated graphically in this figure (e.g., "covarying" vs. "ambiguous-VOT"); however, within a given block, all stimuli were presented randomly, and these different types of stimuli were undifferentiated from the listeners' point of view

and the ambiguous stimuli within each block makes it possible to test how phonetic categorization of the same ambiguous stimuli varies as a function of whether listeners are exposed to canonically vs. reversely covarying stimuli.

Procedure

The experiments took place at the Hanyang Phonetics and Psycholinguistics Laboratory at Hanyang University, Seoul (for native Korean listeners) and at the Auditory Cognitive Neuroscience Laboratory at the University of Arizona, Tucson (for native English listeners). Participants sat in front of a computer in sound-attenuated booths and received both oral and written instructions in English telling them that they would hear English sentences containing either "pa" or "ba" and that they should press 'p' or 'b' to indicate which sound they heard. They were told that the experiment would be divided into five blocks, but were not informed that the blocks would be in any way different from one another. The Neutral block was presented first for all participants. This was followed by two blocks of Canonical and two blocks of Reversed (for half the subjects), or two blocks of Reversed and two blocks of Canonical (for the other half), such that each subject completed one Neutral, two Canonical and two Reversed blocks. Each subject heard 162 trials in the Neutral condition (two randomized repetitions of the baseline stimuli) and 280 trials in each of the Canonical and Reversed conditions (ten randomized repetitions of the covarying-plus-ambiguous stimulus set, times two blocks). The covarying stimuli were not differentiated from the ambiguous stimuli for the participants; all stimuli within a given block were intermixed randomly. The experiment took about 25 min.

Grouping of participants: reliance scores

Based on previous work in which Korean listeners were found to use different cue weighting strategies for the English stop voicing contrast (Schertz et al., 2015), participants were expected to show different patterns of categorization for the covarying stimuli: a "VOT group" classifying stimuli with long VOT as voiceless and short VOT as voiced (irrespective of f0), a "f0 group" classifying stimuli with high f0 as voiceless and low f0 as voiced (irrespective of VOT), and a "VOT+f0" group classifying only stimuli with high f0 and long VOT as voiceless and all other stimuli as voiced (schematic in Fig. 3). All participants were expected to categorize the covarying stimuli in the Canonical block in the same way, perceiving the long VOT, high f0 (Quadrant I) stimuli as voiceless /p/ and the short VOT, low f0 (Quadrant III) stimuli as voiced /b/. However, different patterns were expected in the covarying stimuli in the Reversed condition (Quadrants II and IV), and we therefore used listeners' responses to stimuli in these two quadrants to separate them into groups: (1) *VOT group* (Quadrant II = voiceless and Quadrant IV = voiced); (2) *f0 group* (Quadrant II = voiced and Quadrant IV = voiceless); and (3) *VOT+f0 group*, (Quadrant II and Quadrant IV = voiced).

We calculated a "reliance score" (similar to the "reliance ratio" used by Escudero and Boersma (2004) and Kondaurova

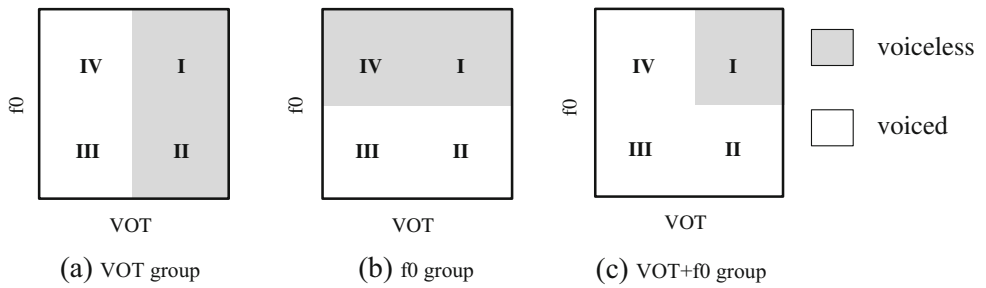


Fig. 3 Schematic of predicted responses for Korean listeners with different primary cue reliance in classifying covarying stimuli (collapsed over Canonical and Reversed blocks)

and Francis (2010) in their examination of spectral vs durational cue weighting in the English /i/-/ / contrast) for each participant by taking the difference between the ratio of "voiceless" response to covarying stimuli in Quadrant II and Quadrant IV. We expected listeners to fall into three groups, with some clustering near 1 (relying exclusively on VOT), some clustering near –1 (relying exclusively on f0), and some clustering around 0 (equal reliance on VOT and f0). Since we predicted that listeners would modify use of their secondary, but not their primary, cue, different adaptation patterns were expected for these different groups; therefore, the subsequent analyses were performed separately for each group.

Statistical analyses

The goal of this work was to assess how listeners adjusted their categorization of the four ambiguous stimuli (i.e., ambiguous-VOT with high or low f0 and ambiguous-f0 with long or short VOT) based on the different distributional information across blocks (i.e., the covarying stimuli in the Canonical vs Reversed blocks), and how these adaptation patterns differed as a function of listeners' initial cue weighting strategies. We quantified the use of each cue (VOT, f0) in each block by taking the difference in "voiceless" responses to the high and low versions of the ambiguous stimuli for that cue. For example, the use of f0 for a block was determined by the difference in "pa" responses to the ambiguous-VOT stimulus with *high* f0 and the ambiguous-VOT stimulus with *low* f0. Adaptation was then defined as a significant change in the difference score for the Reversed block (as determined by a paired-sample *t*-test). Cohen's *d* was used as a measure of effect size.

Predictions

In line with Idemaru and Holt (2011)'s findings using the same paradigm, English listeners were expected to decrease their reliance on f0 in categorization of ambiguous-VOT stimuli when confronted with noncanonical use of f0. Specifically, we expected greater use of f0 (i.e., a larger f0-difference score) in the Canonical block than in the Reversed block. Since VOT is the primary cue to the stop distinction for native English listeners, and the stimuli were chosen specifically to have unambiguous values of VOT, we did not expect to see any change in listeners' use of VOT in classifying the ambiguous-f0 stimuli. Therefore, similar VOT-difference scores for the ambiguous-f0 stimuli were expected in the Canonical and the Reversed blocks for native English listeners.

Our primary questions of interest involve the L1 Korean/L2 English listeners. First, we wanted to test whether they showed adaptation at all. If non-native listeners do adapt, we expected that they might show similar adaptation strategies as native listeners (i.e., that category-internal "dimension-based statistical learning" underlies L2 as well as L1 category tuning). In this case, the Koreans using primarily VOT (i.e., those listeners whose cue-weighting strategies reflect those of native English listeners) should show native-like patterns when classifying the ambiguous stimuli (adaptation of f0 but not of VOT). Based on the hypothesis that the dominant cue serves as an anchor for adaption, the Korean f0 group was predicted to show the opposite pattern (adaptation of VOT but not of f0). These polarized adaptation patterns would be due solely to individual differences in initial cue weighting strategies, as the stimuli presented to each group are identical. Since the Korean VOT+f0 group requires both long VOT and high f0 to classify a stimulus as a voiceless stop, the predictions were less clear for this group. However, since they appeared to have more evenly distributed weights between the two cues than the listeners in the other two groups, the change in VOT-difference scores across the two blocks was expected to be comparable to the change in f0-difference scores across blocks.

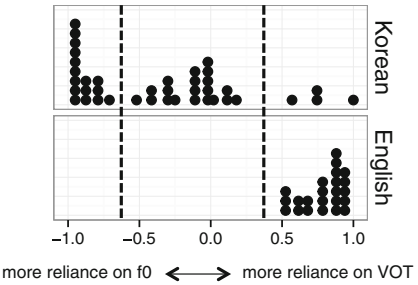


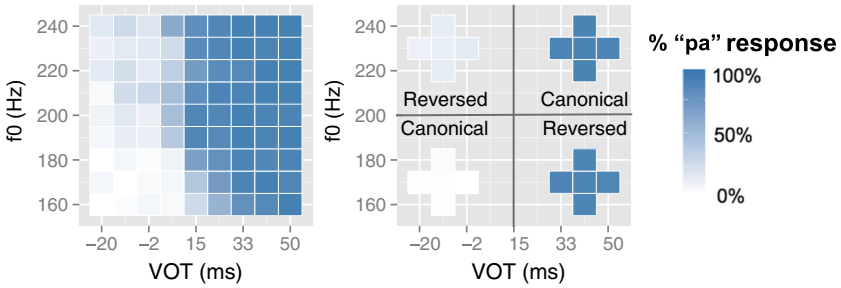
Fig. 4 Reliance differences used for grouping of participants: difference in ratio of "voiceless" categorization between covarying stimuli in the Reversed condition, Quadrant II (long VOT, low f0) and Quadrant IV (short VOT, high f0), as shown in Fig. 2. Each *dot* represents one listener. A reliance difference of –1 represents full reliance on f0 in the categorization of covarying stimuli, a reliance difference of 1 represents full reliance on VOT, and 0 represents equal reliance on both dimensions

Results

Grouping of participants

Reliance scores for English and Korean listeners (ratio of "voiceless" responses in Quadrant II minus "voiceless" responses in Quadrant IV) are shown in Fig. 4. The Korean listeners clustered in three categories, as expected, with one group showing a greater reliance on f0 (n = 16), one group showing a greater reliance on VOT (n = 4), and the rest of the listeners (n = 20) showing a more equal reliance on both.[1] The English listeners clustered together, relying primarily on VOT.

[1] Further investigation of the factors underlying the individual differences in this population make for an interesting topic for future research. The present groupings do not appear to be predictable based on proficiency or amount of experience with English; however, the group of participants used for this study is not sufficiently large, nor is their experience with English sufficiently heterogeneous, to make claims about this relationship.



(a) Baseline stimuli (b) Covarying stimuli

Fig. 5 L1 English control listeners' performance on responses to baseline stimuli (Block 1) and covarying stimuli (collapsed over both Canonical and Reversed blocks). Each cell represents one stimulus, and the darkness of the cell represents the percentage "voiceless" response in a forced-choice task; the *darkest cells* elicited 100 % 'pa' response, while *white cells* elicited 100 % 'ba'

Results: L1 English

L1 English control listeners' responses to the baseline stimuli (Block 1), as well as their responses to the covarying stimuli (pooled across the Canonical and Reversed blocks) are shown in Fig. 5. These patterns indicate that English listeners indeed categorized stimuli based primarily on VOT, classifying stimuli with long VOT as voiceless (despite low f0 in the Reversed block in Quadrant II) and stimuli with short VOT as voiced (despite high f0 in the Reversed block in Quadrant IV). Results for the ambiguous test stimuli, shown in the left panel of Fig. 6, showed the expected categorization pattern: the f0-difference scores for the ambiguous-VOT test stimuli were greater in the Canonical than in the Reversed block [$t(22) = 7.00$, $P < .001$, $d = 1.46$]. This indicates that English listeners made less use of the secondary cue of f0 when exposed to an "accent" showing noncanonical patterning of VOT and f0 (i.e., in the Reversed block). On the other hand, quite a different categorization pattern was observed when listeners categorized ambiguous-f0 stimuli. As can be seen in the right panel of Fig. 6, English listeners relied consistently on VOT

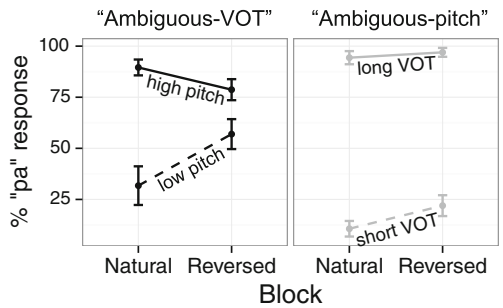


Fig. 6 L1 English control listeners' responses to ambiguous-VOT (left) and ambiguous-f0 (right) stimuli. The *y-axis* shows percentage 'pa' response in a forced-choice ('ba'-'pa') task across blocks. Ambiguous-VOT stimuli with high f0 were classified as mostly 'pa' and those with low f0 as mostly 'ba' (i.e. a large f0-difference score) in the Canonical block; however, this f0-difference score was greatly diminished in the Reversed block, showing a reduced use of f0 in categorization. For ambiguous-f0 stimuli, listeners showed large VOT-difference scores in both blocks (e.g., classified stimuli with short VOT as 'ba' and long VOT as 'pa'), although this effect was slightly smaller in the Reversed block

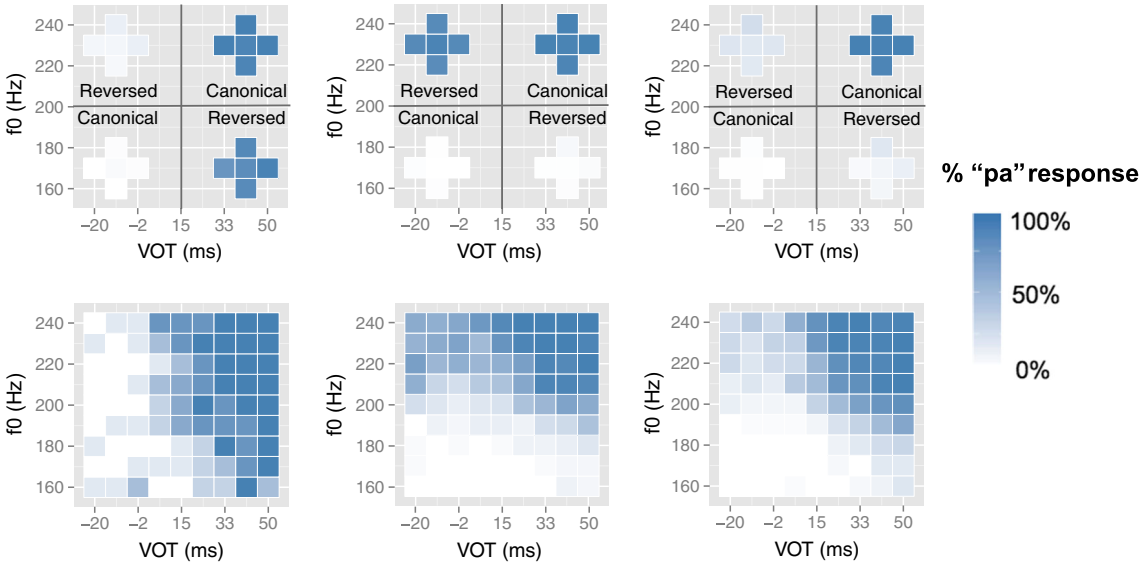
in categorizing ambiguous-f0 stimuli even in the Reversed block. This is again consistent with the prediction that listeners do not use the secondary cue (f0) as an anchor for adaptation, so that even if f0 information is not matched canonically with VOT in the Reversed block, listeners still use VOT as a reliable cue to the voicing contrast. English listeners made slightly less use of VOT in categorizing ambiguous-f0 stimuli in the Reversed block—an unexpected result based on our predictions. That is, the VOT-difference scores for the ambiguous-f0 stimuli was slightly smaller in the Reversed than in the Canonical block [$t(22) = 2.81$, $P < .05$, $d = 0.59$], although the difference across blocks is much smaller than the effect of f0 on categorization of the ambiguous-VOT stimuli.[2]

Results: L1 Korean/L2 English listeners

As discussed above, Korean participants were grouped into those listeners who used primarily VOT, primarily f0, or a combination of the two. Each group's responses to the covarying stimuli (collapsed across the Canonical and Reversed blocks), as well as their baseline cue weights, are shown in Fig. 7, and responses to the ambiguous stimuli are shown in Fig. 8.

VOT group ($n = 4$) As can be seen in the left panel of Fig. 8a, the f0-difference scores on ambiguous-VOT stimuli showed a trend toward significance in the expected direction, with an effect size comparable to that of the L1 English listeners [t(3)

[2] We hypothesized that the anomalous change in the use of VOT across blocks may have been related to the fact that there appeared to be a "voiceless" bias for L1 English listeners in this stimulus set (which had been created based on pilot work with Korean listeners); in particular, the 15 ms "ambiguous VOT" tokens were categorized as "pa" 76 % of the time in the baseline condition. We therefore ran a follow-up study which exactly replicated the current work but used a modified stimulus space centered around English listeners' actual VOT boundary on these stimuli (7 ms). This new group of listeners ($n = 24$, from the same population as the original study) showed the expected results, with f0-difference scores on ambiguous-VOT test stimuli greater in the Canonical than the Reversed block [$t(23) = 6.48$, $P < .001$], but no effect of VOT on the ambiguous-f0 stimuli [$t(23) = 1.34$, $P = .19$].



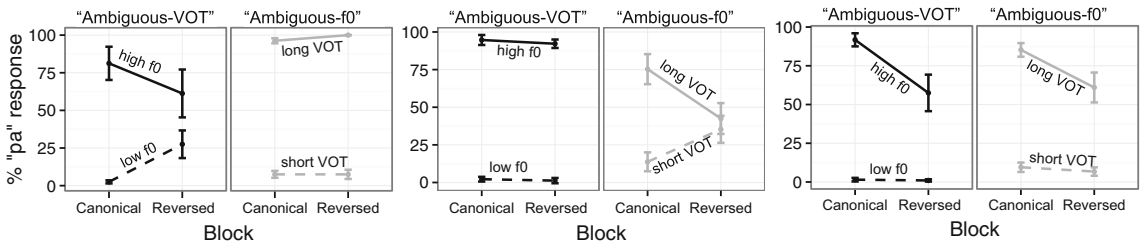
(a) VOT group (n=4) (b) F0 group (n=16) (c) VOT+f0 (n=20)

Fig. 7 Native Korean listeners' responses to covarying stimuli across both the Canonical and Reversed blocks (*top*) and baseline stimuli (*bottom*). The graphs show data averaged across all participants in each group (determined by performance on covarying stimuli, see Fig. 4). Each cell represents one stimulus, and the darkness of the cell represents the percentage "voiceless" response in a forced- choice ('ba'/'pa') task; the *darkest cells* elicited 100 % 'pa' response, while *white cells* elicited 100 % 'ba'

= 2.92, P = .06, d = 1.45]. VOT-difference scores on ambiguous-f0 stimuli were not significantly different between the two blocks [t(3) = −.88, P > .05]. Therefore, although there was limited power given the small number of listeners who relied on VOT initially, the trend suggests that (1) these L2 listeners do show adaptation, and (2) the adaptation is comparable to that of native listeners: specifically, these listeners decreased their reliance on the secondary cue (i.e., f0) in the Reversed block, while showing no modulation of VOT (their primary cue) when classifying ambiguous-f0 stimuli across the two blocks.

F0 group (n = 16) For these listeners, who relied primarily on f0, f0-difference scores on the ambiguous-VOT stimuli were not significantly different between the two blocks [t(15) = 0.52, P > .05] as shown in the left panel of Fig. 8b. This indicates that listeners who used f0 as a primary cue to the contrast did so to an equal extent in the context of both Canonical and Reversed covariation of VOT and f0. On the other hand, as shown in the right panel of Fig. 8b, the VOT-difference scores on the ambiguous-f0 stimuli were significantly greater in the Canonical than in the Reversed block [t(15) = 9.01, P < .001, d = 2.25]. Therefore, these L2 listeners showed a clear change in categorization patterns across blocks, and, as with L1 listeners, this adaptation was characterized by a reduction in secondary cue use in the context of the Reversed block.

VOT+f0 group (n = 20) For listeners in the VOT+f0 group, f0-difference scores were significantly greater in the Canonical than in the Reversed block [t(19) = 5.52, P < .001, d = 1.23] (Fig. 8c, left panel). Similarly, for ambiguous-f0 stimuli, the VOT-difference scores were significantly greater in the Canonical than Reversed block [t(19) = 4.81, P < .001, d = 1.07] (Fig. 8c, right panel). Unlike listeners in the other two groups, the changes in the use of the two cues



(a) VOT Group (b) F0 Group (c) VOT + f0 Group

Fig. 8 L1 Korean/L1 English listeners' responses to a forced-choice task on ambiguous-VOT stimuli with intermediate VOT (*left*) and ambiguous-f0 stimuli with intermediate f0 (*right*), grouped by categorization strategy. The *y-axis* shows percentage 'pa' response in a forced-choice ('ba'-'pa') task, with performance on Canonical vs Reversed blocks shown on the *x-axis* of each panel

were almost identical for these listeners: a within-subjects ANOVA examining the effects of Cue (VOT or f0) and Block (Canonical vs Reversed) showed only an effect for Block [$F(1,19) = 52.43$, $P < .001$], with no effect for Cue [$F(1,19) = 1.47$, $P = .24$] and no interaction between Cue and Block [$F(1,19) = 2.66$, $P = .12$].

Discussion

The performance of the L2 listeners in the current work suggests that sensitivity and rapid adaptability to changes in distributional information across phonetic categories is a hallmark of non-native, as well as native, speech perception. L2 listeners appear to employ similar dimension-based adaptation strategies to those of native listeners, using more reliable phonetic dimensions to extrapolate information about other, secondary, dimensions defining sound categories. Response patterns of L1 English listeners replicated the results of Idemaru and Holt (2011): L1 listeners, who rely primarily on VOT to distinguish English voiced vs voiceless stops, decreased their reliance on f0 (their secondary dimension) when exposed to an "accent" in which VOT and f0 were correlated in a noncanonical direction. On the other hand, Korean listeners who relied primarily on f0 to distinguish their L2 English contrast decreased their reliance on *VOT* (their secondary dimension) when exposed to the same noncanonical accent. As expected, the few Korean listeners who relied primarily on VOT to distinguish the L2 English contrast showed the same trajectory of adaptation as the L1 English listeners did. In both of these cases, while use of the secondary cue decreased in the Reversed block, use of the primary cue (VOT and f0, respectively) appeared to remain stable throughout both blocks, with low values (of VOT or f0) eliciting voiced and high values eliciting voiceless responses.[3] Korean listeners who relied on both f0 and VOT modulated the use of both cues to a comparable extent, an effect that appears to be driven by an overall decrease in "voiceless" responses in these blocks (see below for further discussion on this point).

The rapid adaptation of L2 categories demonstrated here stands in contrast to the lack of plasticity that characterizes the long-term learning of L2 categories (e.g., Han, 2004). While there have been demonstrations of moderate flexibility in L2 categories with extensive training (e.g., Iverson et al., 2005; Kondaurova & Francis, 2010; Escudero et al., 2011; Lim & Holt, 2011), it is surprising that L2 learners would shift category responses in less than 100 exposures to a non-canonical "accent" with no explicit (or lexical) feedback. In fact, the magnitude of L2 category adaptation appears to be comparable to that of native speakers, mirroring results of Reinisch et al. (2013) and Schuhmann (2014). Together, these findings point to strikingly similar adaptation processes across L1 and L2 listeners, at least in the context of phonetic category "tuning" in response to distributional changes (cf. Pajak, Fine, Kleinschmidt, & Jaeger, 2015).

Listeners' variable adaptation patterns were predictable from their initial relative cue weights, in line with work showing differential performance on perceptual learning tasks based on initial listening strategy in both L1 (Sawusch & Nusbaum, 1983) and L2 (Chandrasekaran et al., 2010; Wanrooij et al., 2013) listeners. More specifically, the finding follows straightforwardly from the prediction that listeners use a primary dimension as an anchor to adjust use of a secondary dimension (Idemaru & Holt, 2011), by means of bootstrapping from category-internal distributions of phonetic cues. The direct comparison of groups with different initial cue-weighting strategies highlights the fact that the choice of the anchor dimension depends on the relative weight of the dimensions in listeners' initial definition of the contrast, and that the relative primacy of cues determines the nature of the subsequent category adaptation. In other words, the exact same pattern of distributional changes of phonetic cues in the input may elicit categorically different adaptation strategies, depending solely on variation in listener-specific cue-weighting patterns. One consequence of this result is that even L2 listeners who are similar in accuracy on canonical L2 syllable categorization may have radically different functional categories when confronted with a speaker with a noncanonical accent.

Listeners in the "VOT+f0" group changed the use of both dimensions to an equal extent, lending support to the idea that they do, in fact, weight both dimensions relatively equally. However, the nature of the adaptation differs qualitatively from that of the other two groups. Listeners in the two "unidimensional" (VOT or f0) groups appear to classify both tokens of their ambiguous stimuli at chance in the Reversed block (e.g., in the VOT group, ambiguous-VOT stimuli with both low and high values of f0 are at about 50 %); that is, the listeners in both of these groups appear to actually stop using the secondary dimension as a cue to categorization in the Reversed block (see Fig. 6). On the other hand, the changes seen in the VOT+f0 group can be interpreted more logically as simply an overall decrease in "voiceless" responses in the Reversed block, caused by a shift in category boundary or decision bias rather than a change in cue weighting. If these listeners were actually decreasing their reliance on one or both of the cues, an increase in "voiceless" responses for low values of each cue (relative to their categorization in the Canonical block) as well as a decrease for high values would be expected; however, only the latter was found. One

[3] An anonymous reviewer suggests that listeners' primary cue weights may be increasing (concurrent with a decrease in the secondary cue) during the Reversed blocks—a change that would be undetectable in the current experiment (since use of primary cue was already at ceiling, by design of the paradigm used here). This prediction could be tested in future work with a different paradigm.

explanation for this shift depends on the distribution of stimuli in the Reversed block. Recall that the VOT+f0 group required both long VOT and high f0 to identify a given stimulus as voiceless. At the same time, the Reversed block included co-varying stimuli with either long VOT paired with low f0 or short VOT paired with high f0 (see Fig. 2). Therefore, these listeners were essentially not hearing any good "voiceless" tokens during the Reversed block, which may have caused an overall increase in bias toward choosing "voiced" in these blocks.

The fact that listeners in both the f0 and VOT groups used their primary cue as an anchor from which to bootstrap distributional information about secondary cues from the input demonstrates both the robustness of the primary cues and the flexibility of the secondary cues for each group. The flexibility in the use of VOT by the f0 group is particularly striking. As discussed in the Introduction, there are reasons to expect that VOT might be expected to be an important cue to the L2 English stop contrast, even for those Korean listeners who rely primarily on f0. Most of the native Korean participants in Schertz et al. (2015) showed more reliable differences in VOT than in f0 when *producing* their L2 English stop contrast (including many of the speakers whose primary cue in *perception* was f0). Furthermore, given the overwhelming primacy of VOT in native English productions, it might be expected that even for listeners who have a bias toward relying on f0, their experience with the long-term distributional properties of the stop voicing contrast in English likely demonstrate that VOT is an important cue. The current results, however, show that listeners who rely primarily on f0 can be induced to stop using VOT, highlighting its status as a truly secondary cue. The results also show that for this same group of listeners, f0 on its own is a robust enough cue to underlie distributional learning: even in the absence of prototypical VOT values, stimuli with high f0 are good enough exemplars of voiceless stops (or stimuli with low f0 are good enough exemplars of voiced stops) to be used to anchor learning of secondary cue distributions.

This work focuses on short-term category adaptation to idiosyncratic distributions of sounds; however, some of the questions brought up may extend more broadly to the long-term structure and acquisition of L2 phonetic categories. Many models of categorization assume that cue weights arise from the distributional properties of the input, as approximated by production data (e.g., Nearey, 1997; Nearey & Hogan, 1986; Lotto, Sato, & Diehl, 2004; Toscano & McMurray, 2010), but in non-native listeners, these distributional regularities may be to a large extent masked by native language biases. The current work shows that these two factors cannot be interpreted independently because listener biases interact in a complex way with changes in distributional information. In particular, if it is the case, as proposed above, that listeners decrease their reliance on secondary dimensions when confronted with the sorts of changing distributional patterns used in the present paradigm, then this implies that certain types of short-term distributional variability will actually reinforce initial listener biases in L2 speech perception, even when these initial biases are not the same as those of native listeners, thus potentially in conflict with the long-term distribution of cues in the language. The fact that the same sorts of distributional changes can result in different adaptation patterns needs to be taken into account when considering the contribution of listener biases and distributional regularities in the initial acquisition and ongoing tuning of L2 phonetic categorization.

The current results provide an example of how multiple factors influence how listeners modify their cue weighting strategies (cf. Holt & Lotto, 2006); in particular, the differential trajectories of adaptation, which can be attributed to the same adaptation strategy, highlight the interaction between of statistical learning and initial biases. The rapid response to short-term changes in the input distribution of stimuli could be modeled in an episodic, exemplar-based model (e.g., Goldinger, 1996; Johnson, 1997; Pierrehumbert, 2001). Similarly, "cue-integration" approaches in which distributional information, but not necessarily individual tokens, is stored (HICAT: Smits, 2001a, b, FLMP: Oden & Massaro, 1978, Toscano & McMurray, 2010) would also be able to accommodate the current findings (and these sorts of models can be computationally difficult to separate from exemplar models in terms of categorization, cf. Smits, Sereno, & Jongman, 2006). One other possibility is that the adaptation occurs due to supervised learning as a result of the primary cue activating a phonemic category representation and an error-signal being generated by the mismatch between expected secondary cue relationship for that category and the actual secondary cue input (Guediche et al., 2014). Regardless of the specific model used, the fact that native and non-native listeners demonstrate the same sensitivity and adaptability to changing distributional information in this task suggests that a unified model may be able to account for both L1 and L2 short-term perceptual learning; future work should explore how far this similarity extends to L1 vs. L2 learning more generally (cf. Pajak et al., 2015).

Conclusion

The non-native listeners in the present work made rapid online shifts in their categorization strategies in response to changes in the input by means of category-internal "dimension-based statistical learning" (Idemaru & Holt, 2011), just as native listeners did. The comparison of native Korean/L2 English listeners who used primarily f0 to distinguish the L2 stop voicing contrast with L1 English listeners who use primarily VOT allowed for a direct test of the hypothesis that these modifications result from listeners' choice of one acoustic

dimension as an "anchor" from which to monitor and learn about potentially idiosyncratic use of other dimensions by the current speaker. As predicted, listeners with different anchor dimensions showed categorically different adaptation strategies; in particular, they stopped using their secondary dimension in categorization when the primary and secondary dimensions gave conflicting information about category membership. The current work demonstrates that the individual variability inherent in foreign sound perception can provide a fruitful perspective from which to explore processes underlying more general category learning and adaptation, and the results highlight the fact that models of auditory category learning need to take into account the potential interactions between listeners' initial biases and dynamic adaptation to changes in the current listening environment.

Acknowledgments The authors would like to thank Daejin Kim, Kathy "Nico" Carbonell, Karina Castellanos, and Omar Hussain for their help running subjects, as well as Arthur Samuel, Miquel Simonet, and two anonymous reviewers for helpful feedback on previous versions of this work. This work was supported in part by NSF EAPSI grant #1311026 and NIH-NIDCD grant #R01DC004674.

References

Baese-Berk, M., Bradlow, A., & Wright, B. (2013). Accent-independent adaptation to foreign-accented speech. *The Journal of the Acoustical Society of America, 133*(3), EL174–EL180.

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

Best, C. T. (1995). A direct realist view of cross-language speech perception. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research* (pp. 171–204). Baltimore, MD: York.

Boersma, P., & Weenink, D. (2011). Praat: Doing Phonetics by computer, version 5.3 http://www.praat.org

Bradlow, A., & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition, 106*(2), 707–729.

Chandrasekaran, B., Sampath, P., & Wong, P. (2010). Individual variability in cue-weighting and lexical tone learning. *The Journal of the Acoustical Society of America, 128,* 456–465.

Cho, T., Jun, S.-A., & Ladefoged, P. (2002). Acoustic and aerodynamic correlates of Korean stops and fricatives. *Journal of Phonetics, 30*(2), 193–228.

Clarke, C. M., & Garrett, M. F. (2004). Rapid adaptation to foreign-accented English. *Journal of the Acoustical Society of America, 116,* 3647–3658.

Clayards, M. A., Tanenhaus, M. K., Aslin, R. N., & Jacobs, R. A. (2008). Perception of speech reflects optimal use of probabilistic speech cues. *Cognition, 108*(3), 804–809.

Cutler, A., McQueen, J. M., Butterfield, S., & Norris, D. (2008). Prelexically-driven perceptual retuning of phoneme boundaries. *Proceedings of Interspeech, 2008,* 2056.

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.

Diehl, R. L., Elman, J. L., & McCusker, S. B. (1978). Contrast effects on stop consonant identification. *Journal of Experimental Psychology: Human Perception and Performance, 4*(4), 599–609.

Eimas, P. D., & Corbit, J. D. (1973). Selective adaptation of linguistic feature detectors. *Cognitive Psychology, 4,* 99–109.

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

Escudero, P., Benders, T., & Wanrooij, K. (2011). Enhanced bimodal distributions facilitate the learning of second language vowels. *The Journal of the Acoustical Society of America, 130*(4), EL206–EL212.

Escudero, P., & Boersma, P. (2004). Bridging the gap between L2 speech perception research and phonological theory. *Studies in Second Language Acquisition, 26,* 551–585.

Flege, J. E. (1995). Second language speech learning: Theory, findings, and problems. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research* (pp. 233–277). Baltimore, MD: York.

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

Francis, A., & Nusbaum, H. (2002). Selective attention and the acquisition of new phonetic categories. *Journal of Experimental Psychology: Human Perception and Performance, 28*(2), 349–366.

Goldinger, S. D. (1996). Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*(5), 1166–1183.

Goudbeek, M., Smits, R., Cutler, A., & Swingley, D. (2005). Acquiring auditory and phonetic categories. In H. Cohen & C. Lefebvre (Eds.), *Handbook of categorization in cognitive science* (pp. 497–513). Amsterdam: Elsevier.

Goudbeek, M., Swingley, D., & Smits, R. (2009). Supervised and unsupervised learning of multidimensional acoustic categories. *Journal of Experimental Psychology: Human Perception and Performance, 35*(6), 1913–1933.

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), 1–16.

Han, Z.-H. (2004). *Fossilization in second language acquisition*. Clevedon: Multilingual Matters.

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

Holt, L. L., & Lotto, A. J. (2008). Speech perception within an auditory cognitive neuroscience framework. *Current Directions in Psychological Science, 17*(1), 42–46.

Holt, L., & Lotto, A. (2010). Speech perception as categorization. *Attention, Perception, & Psychophysics, 72*(5), 1218–1227.

House, A. S., & Fairbanks, G. (1953). The influence of consonant environment upon the secondary acoustical characteristics of vowels. *The Journal of the Acoustical Society of America, 25*(1), 105–113.

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

Iverson, P., Hazan, V., & Bannister, K. (2005). Phonetic training with acoustic cue manipulations: A comparison of methods for teaching English /r/-/l/ to Japanese adults. *The Journal of the Acoustical Society of America, 118*(5), 3267–3278.

Iverson, P., Kuhl, P., Akahane-Yamada, R., Diesch, E., Tohkura, Y., Kettermann, A., & Siebert, C. (2003). A perceptual interference account of acquisition difficulties for non-native phonemes. *Cognition, 87*(1), B47–B57.

Johnson, K. (1997). Speech perception without speaker normalization: An exemplar model. In K. Johnson & J. W. Mullennix (Eds.), *Talker variability in speech perception* (pp. 145–166). New York: Academic.

Kingston, J., & Diehl, R. (1994). Phonetic knowledge. *Language, 70*(3), 419–454.

Kingston, J., Diehl, R., Kirk, C., & Castleman, W. (2008). On the initial perceptual structure of distinctive features: The [voice] contrast. *Journal of Phonetics, 36,* 28–54.

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

Kondaurova, M. V., & Francis, A. L. (2010). The role of selective attention in the acquisition of English tense and lax vowels by native Spanish listeners: Comparison of three training methods. *Journal of Phonetics, 38,* 569–587.

Kong, E., & Yoon, I. H. (2013). L2 proficiency effect on the acoustic cue-weighting pattern by Korean L2 learners of English: Production and perception of English stops. *Journal of the Korean Society of Speech Sciences, 5*(4), 81–90.

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

Lee, H., & Jongman, A. (2012). Effects of tone on the three-way laryngeal distinction in Korean: An acoustic and aerodynamic comparison of the Seoul and South Kyungsang dialects. *Journal of the International Phonetic Association, 42*(2), 145–169.

Lim, S.-J., & Holt, L. (2011). Learning foreign sounds in an alien world: Videogame training improves non-native speech categorization. *Cognitive Science, 35*(7), 1390–1405.

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

Llanos, F., Dmitrieva, O., Shultz, A., & Francis, A. L. (2013). Auditory enhancement and second language experience in Spanish and English weighting of secondary voicing cues. *Journal of the Acoustical Society of America, 134*(3), 2213–2224.

Lotto, A. J., Sato, M., & Diehl, R. L. (2004). Mapping the task for the second language learner: The case of Japanese acquisition of /r/ and /l/. In J. Slifka, S. Manuel, & M. Matthies (Eds.), *From sound to sense: 50+ years of discoveries in speech communication* (pp. C381–C386). Cambridge, MA: MIT Press.

Maye, J., Aslin, R. N., & Tanenhaus, M. K. (2008). The weckud wetch of the wast: Lexical adaptation to a novel accent. *Cognitive Science, 32,* 543–562.

Miyawaki, K., Jenkins, J. J., Strange, W., Liberman, A. M., Verbrugge, R., & Fujimura, O. (1975). An effect of linguistic experience: The discrimination of [r] and [l] by native speakers of Japanese and English. *Attention, Perception, & Psychophysics, 18*(5), 331–340.

Moulines, E., & Charpentier, F. (1990). F0-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication, 6,* 453–467.

Nearey, T. (1997). Speech perception as pattern recognition. *Journal of the Acoustical Society of America, 101*(6), 3241–3254.

Nearey, T. M., & Hogan, J. T. (1986). Phonological contrast in experimental phonetics: Relating distributions of production data to perceptual categorization curves. *Experimental Phonology,* 141–161.

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

Oden, G. C., & Massaro, D. W. (1978). Integration of featural information in speech perception. *Psychological Review, 85,* 172–191.

Pajak, B., Fine, A. B., Kleinschmidt, D., & Jaeger, T. F. (2015). Learning additional languages as hierarchical inference: Insights from L1 processing. *Language Learning*, (in press).

Park, H., & de Jong, K. (2008). Perceptual category mapping between English and Korean prevocalic obstruents: Evidence from mapping effects in second language identification skills. *Journal of Phonetics, 36,* 704–723.

Pierrehumbert, J. (2001). Exemplar dynamics: Word frequency, lenition and contrast. In J. Bybee & P. Hopper (Eds.), *Frequency effects and the emergence of linguistic structure* (pp. 323–418). Amsterdam: Benjamins.

Reinisch, E., Weber, A., & Mitterer, H. (2013). Listeners retune phoneme categories across languages. *Journal of Experimental Psychology: Human Perception and Performance, 39*(1), 75–86.

Samuel, A. G., & Kraljic, T. (2009). Perceptual learning for speech. *Attention, Perception, & Psychophysics, 71*(6), 1207–1218.

Sawusch, J. R., & Nusbaum, H. C. (1983). Auditory and phonetic processes in place perception for stops. *Perception & Psychophysics, 34*(6), 560–568.

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

Schuhmann, K. S. (2014). Perceptual learning in second language learners. Ph.D. thesis, Stony Brook University.

Smits, R. (2001a). Evidence for hierarchical categorization of coarticulated phonemes. *Journal of Experimental Psychology: Human Perception and Performance, 27*(5), 1145–1162.

Smits, R. (2001b). Hierarchical categorization of coarticulated phonemes: A theoretical analysis. *Perception & Psychophysics, 63*(7), 1109–1139.

Smits, R., Sereno, J., & Jongman, A. (2006). Categorization of sounds. *Journal of Experimental Psychology: Human Perception and Performance, 32*(3), 733–754.

Sumner, M., & Samuel, A. G. (2009). The effect of experience on the perception and representation of dialect variants. *Journal of Memory and Language, 60*(4), 487–501.

Toscano, J. C., & McMurray, B. (2010). Cue integration with categories: Weighting acoustic cues in speech using unsupervised learning and distributional statistics. *Cognitive Science, 34*(3), 434–464.

Trude, A. M., & Brown-Schmidt, S. (2012). Talker-specific perceptual adaptation during online speech perception. *Language and Cognitive Processes, 27*(7-8), 979–1001.

Wanrooij, K., Escudero, P., & Raijmakers, M. E. J. (2013). What do listeners learn from exposure to a vowel distribution? An analysis of listening strategies in distributional learning. *Journal of Phonetics, 41*(5), 307–319.

Whalen, D. H., Abramson, A. S., Lisker, L., & Mody, M. (1993). F0 gives voicing information even with unambiguous voice onset times. *Journal of the Acoustical Society of America, 93*(4), 2152–2159.

Yamada, R., & Tohkura, Y. (1990). Perception and production of syllable-initial English /r/ and/l/ by native speakers of Japanese. *ICSLP*, pp. 757–760.