Recent findings indicate that listeners are sensitive to talker differences in phonetic properties of speech, including voice-onset-time (VOT) in word-initial stop consonants. The current work extends our earlier research by examining the degree to which listeners adjust the initial mapping from acoustic signal to segmental representation on a talker-specific basis. Two groups of listeners are exposed to a talker producing cane. Word-initial VOTs are manipulated such that one group hears cane produced with short VOTs and the other group hears cane produced with relatively longer VOTs. Following training, listeners’ voicing boundary for a /g/-/k/ continuum is tested. In addition, listeners are tested on phonetic category space by rating members of the continuum for typicality as /k/. If listeners adjust segmental mapping to accommodate talker differences in phonetic properties of speech, then we expect to observe a displacement in the voicing boundaries in line with earlier exposure. Moreover, if this adjustment entails a comprehensive reorganization of phonetic category space, then the /k/ exemplars rated most prototypical will also be displaced for the two listener groups. These data will be discussed in terms of potential constraints on talker-specificity in spoken language processing.
INTRODUCTION

One major goal of research within the domain of speech perception has been to describe how listeners recover the segmental structure of language given that there is no one-to-one mapping between the acoustic signal and individual speech segment (i.e., individual consonants and vowels). Many factors contribute to this variability (e.g., Miller, 1981), including idiosyncratic differences in pronunciation across individual talkers (e.g., Theodore et al., 2009). For many years, variation associated with a talker’s phonetic signature was considered problematic noise for the perceptual system and it was posited that such variation was discarded early in the processing stream. However, this view is now being challenged by a host of findings indicating that listeners encode a wide range of phonetic variation in memory, including that associated with individual talkers’ voices, and that this information can be used to facilitate spoken language processing (e.g., Church & Schacter, 1994; Clarke & Garrett, 2004; Nygaard et al., 1994).

Such investigations have documented the use of talker-specific phonetic information at higher levels of language comprehension such as word recognition, but do not specify precisely which aspects of the acoustic signal are coded on a talker-specific basis and how this information is used to facilitate speech processing. There is a growing body of literature suggesting that the processing benefits observed at higher levels of speech processing might arise through mechanisms that allow listeners to retain talker-specific phonetic detail at the earliest stages of linguistic processing. For example, Norris et al. (2003) demonstrated that when presented with a speech sound that is ambiguous between two segmental categories, listeners use lexical information to modify the boundary between the two categories. Relevant to the current work, this lexically informed boundary adjustment is often applied on a talker-specific basis (Eisner & McQueen, 2005; Kraljic & Samuel, 2005; but see also Kraljic & Samuel, 2007). Thus, one way that listeners accommodate a talker’s idiosyncratic productions, particularly for ambiguous productions, is to customize the initial mapping between the acoustic signal and speech segment by dynamically adjusting perceptual boundaries.

Listeners are also exposed to talker differences in phonetic properties of speech that signal well-defined, unambiguous category members (Newman et al., 2001; Peterson & Barney, 1952) and research has shown that listeners are sensitive to talker differences for individual phonetic properties of speech. For example, Theodore and Miller (2010) examined sensitivity to talker differences in voice-onset-time (VOT), a primary marker of the voicing contrast for word-initial stop consonants. In their study, listeners were exposed to the speech of two female talkers. Characteristic VOTs were manipulated during training phases such that one talker produced /p/ with relatively short VOTs and the other talker produced /p/ with relatively longer VOTs. At test listeners were presented with a short-VOT and a long-VOT variant of /p/ and asked to select which was most representative of a particular talker. The results showed that listeners selected the VOT variant in line with their previous exposure to the talkers’ voices, which indicates that listeners code VOT on a talker-specific basis. Sensitivity to talker differences in VOT raises the possibility that listeners may use this information to customize the mapping between the acoustic signal and phonetic category on a talker-specific basis, as has been shown for the accommodation of ambiguous characteristic productions.

Most research on perceptual learning of talker-specific phonetic detail at this level has evaluated listeners’ ability to shift phonetic category boundaries as a consequence of exposure to a particular talker’s characteristic productions. However, other research, not at the level of individual talkers, has shown that phonetic categories are marked not only by a category boundary, but that they also exhibit a graded internal structure, in that not all members of a particular phonetic category are considered equally good members (Miller & Volaitis, 1989). For example, the VOTs corresponding to best exemplars of a voiceless stop category are robustly influenced by speaking rate, with the best exemplar region located at longer VOTs for a slow compared to a fast speaking rate, an event that mirrors how these values pattern in speech production (Miller & Volaitis, 1989). As part of a larger project, here we test the hypothesis that experience with a talker’s voice fundamentally modifies the category structure of speech sounds in line with that talker’s characteristic productions. That is, we seek to determine whether sensitivity to well-defined exemplars on a talker-specific basis entails not only knowledge of which VOTs are characteristic of a talker (Theodore & Miller, 2010), but also affects listeners’ judgments of category goodness and even the phonetic category boundary across a range of VOTs.

In the current work, two groups of listeners were exposed to the speech of two female talkers, fictitiously referred to as Sheila and Joanne. One group of listeners heard Joanne produced /k/ with relatively short VOTs and Sheila produce /k/ with relatively long VOTs. The other group of listeners heard the opposite pattern of VOT exposure. Testing for both training groups was identical and consisted of three tasks using Joanne’s voice: (1) a
two-alternative, forced-choice (2AFC) task where listeners were presented with a short-VOT and a long-VOT variant of *cane* and asked to choose which is more representative of Joanne’s voice; (2) a category identification task where members of a VOT continuum from *gain* to *cane* were presented to listeners to identify the initial consonant; and (3) a goodness rating task, where listeners rated the goodness of members of a VOT continuum that spanned the voiceless VOT region presented during training as exemplars of Joanne’s voice. Based on previous research (Theodore & Miller, 2010), we predicted that results of the 2AFC test would show that listeners would select the variant consistent with the VOTs heard during training, indicating that listeners are sensitive to talker differences in VOT. The primary question is whether such sensitivity will promote a fundamental talker-specific remapping between the acoustic signal and segmental category. If so, then we predict that the category identification and goodness rating data will reveal a shift in the voicing boundary as well as the internal structure of the voiceless category. An alternative is that given the unambiguous nature of the productions to be learned, accommodation of talker-specific phonetic variation will be reflected solely in the internal category structure without a concomitant change in the voicing boundary.

**METHODS**

**Participants**

To date, 13 adults have participated in this study. Six were assigned to the J-SHORT/S-LONG training group and seven were assigned to the J-LONG/S-SHORT training group. As described below, the difference between the training groups was characteristic VOT production of the two talkers. All participants were native speakers of American English, were not fluent in any other language, had no history of speech/language disorders, and passed a pure-tone hearing screening on the day of testing. All participants were paid $15 USD for their participation.

**Stimuli**

The stimuli for this experiment consisted of a two synthesized continua ranging from *gain* to *cane*, each produced by one of two talkers fictitiously labeled “Sheila” and “Joanne.” Stimulus preparation followed the procedure outlined in Theodore and Miller (2010). In brief, two female native speakers of English with perceptually distinct voices were recorded producing many repetitions of *gain* along with many filler items. One repetition of *gain* was selected for each talker. The selected repetitions had equal VOTs and were equated for word duration and root-mean-square (RMS) amplitude. A pitch-synchronous LPC analysis was performed on each selected token. Using the results of this analysis, two continua were created (one for each talker) by manipulating parameters of the LPC analysis and synthesizing additional tokens based on the new parameters. This procedure yielded, for each talker, a series of tokens that perceptually ranged from *gain* to *cane*. Acoustically, successive steps on the continuum differed in VOTs by approximately 4 milliseconds.

Subsets of these two continua were used during training and test phases. Two sets of training stimuli were created, one for each training group. All training sets consisted of the *gain* endpoint tokens, two tokens from the short-VOT *cane* region, and two tokens from the long-VOT *cane* region. For the J-SHORT/S-LONG training group, the short-VOT *cane* tokens were drawn from Joanne’s continuum and the long-VOT *cane* tokens were drawn from Sheila’s continuum. For the J-LONG/S-SHORT training group, the long-VOT tokens were drawn from Joanne’s continuum and the short-VOT tokens were drawn from Sheila’s continuum. Two amplitude variants of each selected token corresponding to RMS amplitude of the short-VOT and long-VOT tokens were created in order to eliminate a potential amplitude-based confound. In both sets of training stimuli, VOT for the *gain* tokens was 25 ms, the short-VOT *cane* tokens ranged in VOT from 78-87 ms, and VOTs for the long-VOT *cane* tokens range from 170-179 ms.

Three sets of test stimuli were assembled, one for each test task. Testing stimuli were the same for both training groups and consisted of tokens drawn from Joanne’s *gain-cane* continuum. For the 2AFC test, stimuli consisted of four pairs of stimuli. Each pair contained a short-VOT and long-VOT variant of *cane* that was not used during training. Half of the pairs presented the short-VOT variant first with the other half presented the long-VOT variant first. For the identification task, 12 consecutive members of Joanne’s *gain-cane* continuum were selected ranging in word-initial VOTs from 25-74 ms in approximately 4 ms steps. Perceptually, the selected tokens ranged from a clear *gain* to a clear *cane*, with some ambiguous tokens intermediate to the endpoints. For the goodness rating task, a different 12 members of Joanne’s *gain-cane* continuum were selected. These members ranged in VOT from 74-
174 ms in approximately 8 ms steps. Relative to the stimuli presented during training, test stimuli for the 2AFC test represent only the short-VOT and long-VOT variants heard during training. VOTs for the identification test represent the range of VOTs from the voiced to the short-VOT voiceless region of the training stimuli. The stimuli used for the goodness rating test have VOTs that span the short-VOT and long-VOT voiceless regions presented during training. Thus, the three sets of test stimuli appropriately sample different regions of acoustic-phonetic space in line with the hypothesis examined in each test case.

**Procedure**

Listeners completed all testing individually in a sound-attenuated booth and were seated at a small table that contained a computer monitor and a response box. Visual feedback during training was presented on the computer monitor. Auditory stimuli were presented binaurally over headphones at a comfortable listening level that was consistent across listeners.

The overall procedure required listeners to participate in familiarization, training, and testing phases. The goal of familiarization was to give each listener a chance to learn the names of the talkers. During the familiarization phase, two randomizations of the training stimuli appropriate for a given training group were presented. On each trial, the name of the talker appeared on the computer screen while the auditory stimulus was presented. Listeners were instructed to listen to each word and learn the talkers’ names. No responses were collected during the familiarization phase. During training phases, two randomizations of the appropriate training stimuli were presented. On each trial, listeners were instructed to indicate the talker and the initial consonant using one of four buttons labeled “SG,” “SK,” “JG,” and “JK.” Feedback was provided only for the talker decision.

The procedure during test phases differed depending on the particular test, though listeners were told that they would be tested on Joanne’s voice in all cases. No feedback was provided in any of the test phases. For the 2AFC test, two randomizations of the four test pairs were presented. Listeners were directed to indicate which member of the pair was most like Joanne by pressing a button labeled “1” for the first member of the pair or a button labeled “2” for the second member of the pair. For the identification test, one randomization of the 12 identification test tokens was presented and listeners were instructed to indicate whether the initial consonant in each trial was “G” or “K” by pressing the appropriately labeled button. For the goodness rating test, one randomization of the 12 goodness rating test tokens was presented. Listeners were instructed to rate each stimulus in terms of its goodness for Joanne’s voice on a scale of 1-7 by pressing the appropriate button on the response box, with 7 indicating the best exemplar.

Following the familiarization phase, listeners completed three runs that consisted of eight alternating rounds of training and test phases. The three runs were blocked by test task, with all listeners first completing the 2AFC task, and the order of the identification and goodness rating tasks counter-balanced within each training group. The entire experiment lasted approximately 90 minutes.

**RESULTS**

**Training**

For each participant, accuracy during the training phases was assessed separately for the phonetic decision and the talker decision. Mean performance across the participants was near ceiling for both the phonetic decision (96%) and talker decision (95%). These results confirm that the voices were perceptually distinct and that our VOT manipulation yielded the intended phonetic contrast.

**Test 1: 2AFC**

Performance for the 2AFC test was analyzed separately for each training group. For each subject, mean percent long-VOT responses was calculated by collapsing over the eight test trials in each test phase and then over the eight test phases. Recall that on a given test trial, listeners could choose either the short-VOT variant or the long-VOT variant of cane; because these responses must sum to 100, we quantified performance solely in terms of percent long-VOT responses in keeping with established convention. Figure 1 shows mean percent long-VOT responses for the two training groups. As predicted, listeners in the J-SHORT/S-LONG training group chose far fewer long-VOT responses (24%) compared to listeners in the J-LONG/S-SHORT training group (77%), indicating that listeners used their experience with Joanne’s voice during training to guide their performance at test. This pattern was confirmed
statistically as follows. Performance in the J-SHORT/S-LONG training group and the J-LONG/S-SHORT training group was significantly different from chance, [t(5) = -3.227, p = .023] and [t(6) = 4.569, p = .004], respectively. Critically, percent long-VOT responses was significantly lower in the J-SHORT/S-LONG training group compared to the J-LONG/S-SHORT training group [t(11) = -5.3965, p < .001].

**FIGURE 1.** Mean percent long-VOT responses for the two training groups. Error bars indicate standard error of the mean.

*Test 2: Identification*

Performance for the identification test phases was analyzed separately for each training group. For each subject, mean percent “K” responses was calculated for each of the 12 test stimuli by collapsing across the eight test phases. Figure 2, panel (a) shows the identification functions for each training group, and inspection of this figure suggests that the voicing boundary did not differ between the two training groups. To examine the statistical significance of this pattern, probit analyses were used to determine the voicing boundary for each subject by fitting an ogive function to the identification responses for each listener, with the mean of the ogive function serving as the metric of the category boundary. The mean boundary for the J-SHORT/S-LONG training group was 58 ms and the mean boundary for the J-LONG/S-SHORT training group was 57 ms. This difference was not statistically reliable [t(11) = 0.277, p = .787].

*Test 3: Goodness ratings*

Performance for the goodness rating test was analyzed separately for each training group. For each subject, mean goodness rating was calculated for each of the 12 test stimuli by collapsing across the eight test phases. Figure 2, panel (b) shows the mean goodness ratings for each training group. Inspection of this figure suggests that the tokens rated as best exemplars for Joanne’s voice differed as a function of training group. For example, consider goodness ratings for the J-SHORT/S-LONG training group. The peak of this function is located at the shortest VOT values, with mean goodness decreasing as VOT increases. In contrast, the peak of the function for the J-LONG/S-SHORT training group is located at the longest VOT values, with lower goodness ratings for shorter VOTs. Such a pattern indicates that experience with Joanne’s voice during the training phases influenced perceived goodness of the VOT continuum. In order to test the statistical significance of this pattern, we quantified performance using standard conventions as outlined in Allen and Miller (2001). In brief, a best exemplar range was calculated for each subject, defined as the range of VOTs that were rated at or above 90% of the highest rating provided by that subject. Both the lower bound [t(11) = -4.616, p < .001] and upper bound [t(11) = -37.470, p = .013] of the best exemplar range were significantly shorter for the J-SHORT/S-LONG training group compared to the J-LONG/S-SHORT training group, confirming that the peaks of the goodness functions were located at a different range of VOTs across training groups.
CONCLUSIONS

The current data contribute to a growing body of literature indicating that listeners adapt to talker-specific phonetic variation very early in the processing stream by customizing the mapping between the acoustic signal and speech segment on a talker-by-talker basis. Moreover, the current results suggest that the mechanisms that listeners use to achieve this customization may depend on the nature of a talker’s characteristic productions. Earlier work has shown that when listeners must incorporate an ambiguous production into an established phonetic category, they do so by modifying the phonetic boundary between speech segments (e.g., Norris et al., 2003). The preliminary results reported here suggest that listeners also customize the mapping process in line with a talker’s clearly defined, unambiguous characteristic productions of voiceless stop consonants, but that they do so by modifying the internal structure of the phonetic category without adjusting the voicing boundary. Collectively, these findings suggest that the precise way in which listeners accommodate talker-specific phonetic variation may be influenced by the nature of a talker’s characteristic productions along an acoustic-phonetic dimension. That is, when the characteristic production is ambiguous between established phonetic categories, listeners adjust the phonetic boundary. In contrast, when a talker’s characteristic productions are readily mapped onto an established phonetic category, customization for that talker may be obtained by shifting the internal structure of that category to be centered on the characteristic productions.

It is important to note that the failure to observe a reliable boundary shift may reflect the low number of participants reported in this experiment, and additional data are currently being collected to address this possibility. Moreover, a more complete picture of the degree to which the nature of what is to be learned determines the perceptual adjustment would entail testing the hypothesis that accommodating ambiguous talker-specific productions entails adjustments solely at a phonetic category boundary, and does not extend throughout the internal category structure. Future research is aimed at addressing this question.

REFERENCES


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


