Talker-specific influences on phonetic category structure[a]

Rachel M. Theodore,[b] Emily B. Myers, and Janice A. Lombao
Department of Speech, Language, and Hearing Sciences, University of Connecticut, 850 Bolton Road, Unit 1085, Storrs, Connecticut 06269-1085, USA

(Received 6 February 2015; revised 8 July 2015; accepted 14 July 2015; published online 21 August 2015)

A primary goal for models of speech perception is to describe how listeners achieve reliable comprehension given a lack of invariance between the acoustic signal and individual speech sounds. For example, individual talkers differ in how they implement phonetic properties of speech. Research suggests that listeners attain perceptual constancy by processing acoustic variation categorically while maintaining graded internal category structure. Moreover, listeners will use lexical information to modify category boundaries to learn to interpret a talker's ambiguous productions. The current work examines perceptual learning for talker differences that signal well-defined, unambiguous category members. Speech synthesis techniques were used to differentially manipulate talkers' characteristic productions of the stop voicing contrast for two groups of listeners. Following exposure to the talkers, internal category structure and category boundary were examined. The results showed that listeners dynamically adjusted internal category structure to be centered on experience with the talker's voice, but the category boundary remained fixed. These patterns were observed for words presented during training as well as novel lexical items. These findings point to input-driven constraints on functional plasticity within the language architecture, which may help to explain how listeners maintain stability of linguistic knowledge while simultaneously demonstrating flexibility for phonetic representations. © *2015 Acoustical Society of America*.

[http://dx.doi.org/10.1121/1.4927489]

[TCB] Pages: 1068–1078

I. INTRODUCTION

One hallmark of human cognition is the ability to recognize physically different events in the environment as members of a single cognitive category. Within the domain of speech perception, this ability has been examined with respect to the mechanisms that allow listeners to consistently map the acoustic signal to speech sound categories given that the acoustic information produced for a given consonant or vowel varies each time it is spoken. One source of variability concerns differences in speech production across individual talkers. Talker differences have been observed for indexical variation, including fundamental frequency (e.g., Klatt and Klatt, 1990) and voice quality (Fant, 1993; Murray and Arnott, 1993). Talker differences have also been observed for phonetic properties of speech, which are aspects of the signal that listeners use to recover linguistic meaning. For example, talkers differ in formant frequencies specifying vowels (Peterson and Barney, 1952) and centroid frequency specifying fricatives (Newman *et al.*, 2001), which may reflect physical differences among talkers. Talkers also show idiosyncratic differences in producing phonetic properties of speech including differences in voice-onset-time (VOT) specifying stop consonants (Theodore *et al.*, 2009). Despite the lack of invariance between the acoustic signal and linguistic representation, listeners accurately perceive speech sounds when confronted with talker variation (e.g., Nygaard *et al.*, 1994).

Previous research suggests that listeners achieve stable perception, at least in part, by translating continuous acoustic-phonetic variation into discrete linguistic categories (Cooper *et al.*, 1952). For example, consider the acoustic-phonetic property of VOT. VOT is an articulatory property of stop consonants (e.g., /g/ and /k/) that reflects the time between the release of the complete occlusion necessary for stop consonant production and subsequent onset of vocal fold vibration. In speech production, VOTs for English voiced stops (/b/, /d/, /g/) are generally very short, and VOTs for English voiceless stops (/p/, /t/, /k/) are relatively longer (Lisker and Abramson, 1964). Given a range of acoustic-phonetic variation, such as the range of VOTs specifying word-initial stop consonants, listeners' perception is not linearly related to VOT duration. Rather, it is categorical, with some VOTs identified as voiced stops, a different range of VOTs identified as voiceless stops, and an abrupt discontinuity between the two ranges (Cooper *et al.*, 1952). In other words, listeners appear to impose a perceptual boundary at some particular VOT to mark the voicing contrast. However, findings from other paradigms have shown that perception of speech sounds within a given category, such as voiceless stop consonant, is not all-or-nothing. Rather, speech sound categories have a graded internal structure and are thus organized like other cognitive/perceptual categories, with some category members considered better exemplars than others (Miller, 1994).

[a] Pilot data were presented at the 35th annual conference of the Cognitive Science Society, Berlin, Germany, July 2013; Interspeech 2013, Lyon, France, August 2013; and at the 165th meeting of the Acoustical Society of America, Montréal, Québec, June 2013.

[b] Electronic mail: rachel.theodore@uconn.edu

There is a wide body of evidence demonstrating that speech sound categories, both in terms of category boundaries and internal category structure, remain functionally plastic even in adulthood such that representations are dynamically adjusted in light of systematic acoustic-phonetic variation. For example, in speech production, VOTs systematically increase as speaking rate slows (Miller *et al.*, 1984). Listeners accommodate this contextual influence by shifting both the voicing boundary as well as the range of tokens that are judged the "best" exemplars of the category toward longer VOTs for a slow compared to a fast speaking rate (Volaitis and Miller, 1992). These mechanisms may also underlie listeners' ability to accommodate talker-specific phonetic detail. Indeed, previous research has shown that listeners are sensitive to talker differences in VOT such that they can learn that one talker produces characteristically short VOTs and a different talker produces relatively longer VOTs (Theodore and Miller, 2010). Listener sensitivity to talker differences for individual phonetic properties of speech is a logical precursor to their ability to customize the mapping between the acoustic signal and speech sound for individual talkers.

Indeed, the literature on perceptual learning in speech has demonstrated that listeners can use lexical information to modify category boundaries in light of ambiguity in the acoustic signal (e.g., Norris *et al.*, 2003; Eisner and McQueen, 2005; Kraljic and Samuel, 2005; Kraljic and Samuel, 2007). In this paradigm, listeners are presented with an ambiguous speech sound (such as a fricative midway between /s/ and /ʃ/) during a training phase in which they complete a lexical decision task. The critical manipulation is that lexical information is used to differentially bias listeners' perception of the ambiguous sound. For example, one group might hear the ambiguous sound in words such as *pencil*, where the bias is to perceive it as /s/ and the other group might hear the sound in words such as *ambition*, where the bias is to perceive it as /ʃ/. Following the lexical decision training task, listeners are presented with a non-word to non-word continuum from /asi/ to /aʃi/ and are asked to identify each member as one of those two categories. Results have shown that listeners use lexical information to adjust the category boundary such that the previously ambiguous sound is now incorporated into a segmental category (e.g., Norris *et al.*, 2003). Findings in this domain have shown that these adjustments are often applied on a talker-specific basis (Kraljic and Samuel, 2007), and that they are applied conservatively in that listeners do not adjust the category boundary when the ambiguity can be attributed to an incidental event, such as when viewing a speaker with a pen in her mouth (Kraljic *et al.*, 2008).

To date, the literature on perceptual learning has exclusively focused on how listeners modify perceptual representations to accommodate ambiguous productions, with evidence of learning being measured only with respect to category boundaries (although see Sumner, 2011 for results of exposure which includes unambiguous as well as ambiguous tokens). Given that sources of acoustic-phonetic variability, such as a talker's phonetic signature, more often represent well-defined category members (e.g., Newman *et al.*, 2001; Theodore *et al.*, 2009), a complete account of perceptual learning for speech must consider how listeners adjust speech sound representation for unambiguous members of phonetic categories. Moreover, a complete account of perceptual learning should examine changes in organization that may happen within the category proper, and not focus exclusively on category boundaries. In particular, any evidence that leads the listener to suspect an altered shape to the phonetic category may result in a wholesale shift in the listener's phonetic category structure, constituting changes in the location of the phonetic category boundary as well as changes in the perceived goodness of tokens within the category itself. Alternatively, listeners may instead require more compelling evidence of a movement in the category boundary, and may only alter the location of that boundary when confronted with near-boundary tokens.

Toward this end, the current work examines perceptual learning of talker-specific phonetic detail, focusing on talker-differences in VOT for word-initial stop consonants. Two groups of listeners were exposed to the speech of two talkers. Speech synthesis techniques were used to manipulate the talkers' productions in order to provide differential patterns of characteristic /k/ productions to the two groups of listeners. Following training, we examined potential influences on perceptual organization with respect to both the category boundary and the internal category structure. Experiment 1 examined talker-specific influences on phonetic categories holding the lexical items constant between training and test. In Experiment 2, we examined generalization of learning by training listeners on one set of words and testing on novel lexical items. We first present methods and results for each experiment, and then consider their implications jointly in Sec. IV.

II. EXPERIMENT 1

In Experiment 1, two groups of listeners were exposed to the speech of two female talkers. During training phases, listeners heard "Joanne" and "Sheila" produce tokens of *gain* and *cane*. The acoustic characteristics of the *cane* tokens were manipulated such that one group heard Joanne produce *cane* with short VOTs and Sheila produce *cane* with relatively longer VOTs. The other group of listeners heard the opposite pattern of characteristic VOTs; Joanne produced *cane* with long VOTs and Sheila produced *cane* with relatively shorter VOTs. Critically, all VOT variants presented during training fell within the standard range of VOTs for voiceless stops, and thus were unambiguous productions. All listeners were tested on Joanne's speech in three ways. In order to assess the degree to which listeners encode talker-specific details of speech (see Theodore and Miller, 2010), listeners were given a short-VOT and a long-VOT variant of *cane* and asked to choose which was most representative of Joanne's voice. Moreover, to assess the degree to which this sensitivity affects internal category structure, listeners heard a VOT continuum from *gain* to *cane* and were asked to rate each item for goodness as /k/. In addition, to measure adjustments to the phonetic category boundary, listeners heard the same VOT continuum and were asked to identify each member as beginning with either /g/ or /k/.

Based on previous findings indicating that listeners can track talkers' characteristic VOTs (Theodore and Miller, 2010), we predicted that when presented with two VOT variants of *cane* and asked to indicate which was more representative of Joanne, listeners would choose the VOT variant in line with their previous exposure to Joanne's voice. That is, listeners who heard Joanne produce short VOTs during training would choose the short VOT variant more often than those who heard Joanne produce long VOTs during training. If talker-specific phonetic detail has the same influence on internal category structure as other contextual influences such as speaking rate (e.g., Volaitis and Miller, 1992), then we predicted that category goodness ratings would also pattern in line with exposure during training. Moreover, if accommodating talker-specific productions for well-defined, unambiguous category members results in the same perceptual learning as has been shown for ambiguous productions, then we predicted that the category boundary between /g/ and /k/ will be displaced between the two training groups given the differential exposure to Joanne's characteristic VOTs.

A. Methods

1. Participants

Fifty adults between the ages of 19 and 34 were recruited from the University of Connecticut community to participate in the experiment. Participants were randomly assigned to either the J-SHORT or J-LONG training group (described in detail below) and all were paid for their participation. All participants were native, monolingual speakers of American English with no history of speech, language, or hearing disorders according to self-report. All participants passed a pure-tone hearing screening on the day of testing, administered at 20 dB for octave frequencies between 500 and 4000 Hz. The sample size and stopping rule were determined based on sample sizes that have shown sufficient power to detect similar effects using paradigms equivalent to the ones used here (e.g., Volaitis and Miller, 1992; Theodore and Miller, 2010). Two participants were excluded: one due to an inability to derive a category boundary using the methods outlined below, and one due to an inability to calculate a best exemplar region using the methods outlined below. Of the remaining 48 listeners, 25 participated in the J-SHORT training group and 23 participated in the J-LONG training group.

2. Stimuli

The stimuli consisted of two synthesized VOT continua, each perceptually ranging from *gain* to *cane*. The synthesis procedures followed those outlined in Allen and Miller (2004) and Theodore and Miller (2010), and the stimuli used here were drawn from the continua used in Theodore and Miller (2010). Specifically, a naturally produced token of *gain* was acquired from two female speakers who had perceptually distinct voices. We refer to our speakers fictitiously as Joanne and Sheila. The selected tokens were equated for word duration (568 ms) by deleting energy from the word offset and were then equated for root-mean-square (rms) amplitude. To create each continuum, each *gain* token was first analyzed using an Linear Predictive Coding (LPC)-based synthesizer (ASL, KayPENTAX, Montvale, NJ), which calculated values for numerous parameters of the acoustic signal on a frame-by-frame basis, with each frame corresponding to one cycle of vocal fold vibration. The first step of the continuum was generated by synthesizing a token based on the original LPC analysis. Additional tokens were created by systematically manipulating parameters of the LPC analysis for successive frames in order to change the periodic source to a noise source, each time synthesizing a new token with systematically longer VOT. This procedure yielded, for each talker, 36 tokens that ranged in VOT from approximately 20 to 185 ms, in 4–5 ms steps. Perceptually, each continuum ranged from a clear *gain* to a clear *cane*, with some tokens ambiguous between the two perceptual endpoints, and with many tokens servicing as unambiguous exemplars of *cane*.

Subsets of these continua were selected for use during training and test phases. For training, we selected the following from each talker: one *gain* token, two *cane* tokens with short VOTs that were two steps apart on the continuum, and two *cane* tokens with relatively longer VOTs that were also two steps apart. These tokens were organized into two sets, one for each training group. The J-SHORT training group used Joanne's short-VOT *cane* tokens, Sheila's long-VOT *cane* tokens, and the *gain* tokens from both speakers. The J-LONG training group used Joanne's long-VOT *cane* tokens, Sheila's short-VOT *cane* tokens, and the *gain* tokens for both speakers. VOTs of the training tokens are shown in Table I. Within each training set, we duplicated the *gain* token so as to have equal numbers of *gain* and *cane* items in each set. In addition, we created two amplitude variants for each selected token, corresponding to the rms amplitude of the short- and long-VOT variants, respectively. Thus, each set of training stimuli consisted of 16 tokens.

TABLE I. VOT values (ms) of the *gain* and *cane* training stimuli used in Experiment 1.

Training Group: J-SHORT			
		cane	
Talker	*gain*	Token 1	Token 2
Joanne	22	78	88
Sheila	20	172	181
Training Group: J-LONG			
		cane	
Talker	*gain*	Token 1	Token 2
Joanne	22	170	179
Sheila	20	79	88

Two sets of test stimuli were created, one for use during a two-alternative, forced-choice explicit memory test and one for use during goodness rating and category identification tests. All tests were performed using stimuli from Joanne's continuum only and VOTs for the test stimuli are shown in Table II. For the explicit memory test, a short- and a long-VOT variant of *cane* were selected. Recall that for the training tokens, the selected short- and long-VOT variants were each two steps apart on the continuum; the intermediate tokens were used for the explicit memory test. Two amplitude variants of the selected tokens were created corresponding to mean rms amplitude of the selected short- and long-VOT variants of *cane* used during training. Using the selected explicit memory test tokens, pairs of stimuli were created by concatenating a short- and long-VOT variant, separated by 750 ms of silence. Four test pairs were created with this procedure, half that began with the short-VOT token and half that began with the long-VOT token, with amplitude held constant for a given pair.

For the goodness rating and category identification tests, 24 tokens from Joanne's *gain–cane* continuum were selected that spanned the VOTs presented during training. The first 12 tokens represented 12 successive steps on the continuum beginning with the second step. The other 12 tokens were each 2 steps apart on the continuum. With this procedure, the range of VOTs presented during training was assessed for both the goodness and identification tests, without presenting the exact physical token at both training and test.

TABLE II. VOT values (ms) of the *gain–cane* test continuum used in Experiment 1 and the *goal–coal* test continuum used in Experiment 2. The tokens in bold indicate those used to create the pairs for the explicit memory test.

Step	Experiment 1	Experiment 2
1	25	26
2	30	32
3	33	36
4	39	41
5	43	45
6	47	48
7	51	53
8	56	58
9	60	61
10	65	66
11	69	70
12	74	74
13	**83**	**84**
14	92	92
15	101	101
16	110	110
17	120	118
18	129	128
19	138	137
20	147	145
21	156	153
22	166	163
23	**174**	**176**
24	183	184

3. Procedure

Participants completed the experiment individually in a sound-attenuated booth. All were seated at a table with a computer monitor and a response box. Auditory stimuli were presented via headphones (Sony MDR-V6, Tokyo, Japan) and visual stimuli were displayed on the monitor. Participants completed three cycles of training and test phases, one for each of the three test tasks (explicit memory, goodness rating, and category identification). Procedural details for the training and test tasks are described below. The overall procedure required listeners to alternate between training and test phases in order to help ensure that what was measured during test reflected exposure during training, and not exposure to the test stimuli themselves. Listeners completed six alternations between training and test for each test type (i.e., six training and six test sessions for each type of test task). For example, some listeners first completed six alternations of training and explicit memory test, then six alternations of training and goodness rating test, and finally six alternations of training and category identification test. All listeners completed the explicit memory test task first, and order of the goodness rating and identification tests was counterbalanced across listeners. Prior to the beginning of the experiment proper, listeners participated in a short familiarization phase in order to learn Joanne and Sheila's voices. During familiarization, one randomization of the training stimuli was presented and the name of the talker for each stimulus simultaneously appeared on the computer monitor. Listeners were instructed to listen and learn the names of the talkers; no responses were collected. Listeners also completed a brief practice prior to the first test phase for each type of test. The entire procedure lasted approximately 2 h.

a. Training. During each training phase, 3 randomizations of the 16 training stimuli were presented. On each trial, participants were asked to identify the initial consonant and talker by pressing an appropriate button on the response box. Feedback was provided for talker choice only. Feedback appeared on the monitor for 1500 ms following each response, and the next trial began 2000 ms after the offset of visual feedback. Thus, during the training phases listeners learned how Joanne and Sheila produced the words *gain* and *cane*; critically, we manipulated each talker's characteristic VOTs between the two training groups.

b. Test. Though listeners were exposed to both Joanne and Sheila's voices during training, we tested all listeners using Joanne's voice only across three types of test phases. For all three tests, listeners were directed to make their decisions based on experience with Joanne's voice during training in order to increase attention to the salient acoustic characteristics. In each explicit memory test phase, participants were presented with two randomizations of the four explicit memory test pairs and were asked to select which member of the pair was most representative of Joanne's voice. They pressed a button labeled "1" to indicate the first member of the pair or a button labeled "2" to indicate the second member of the pair. No feedback was provided and the next trial began 2000 ms after each response.

In each goodness rating test phase, listeners heard one randomization of the 24-member continuum and were asked to rate each member for goodness as /k/ by pressing an appropriately labeled button. They made their decision on a scale from 1 to 7, with 7 being most representative of their previous experience with Joanne's voice. In each category identification test phase, listeners also heard one randomization of the 24-member continuum, but they were asked to identify the initial consonant of each item as either /g/ or /k/ by pressing an appropriately labeled button. No feedback was provided for either the goodness or identification test phases. The inter-trial interval was 2000 ms for both test phases, timed from each response to the onset of the next auditory stimulus.

B. Results

1. Training

Performance during training was analyzed by calculating percent correct for the phonetic and talker decisions separately for Joanne and Sheila's voices. Performance for both voices reached ceiling during the first six training phases, for both decisions. Mean accuracy for the phonetic decision was 97.07% [standard deviation (SD) = 6.80] for Joanne's voice and 92.14% (SD = 12.89) for Sheila's voice. Mean accuracy for the talker decision was also high, 95.86% (SD = 5.73) and 94.59% (SD = 6.10) for Joanne and Sheila's voices, respectively.

2. Test

a. Explicit memory. Performance for the explicit memory test phases was analyzed by calculating percent long-VOT responses across the six explicit memory test phases. Mean percent long-VOT responses for the J-SHORT training group was 22.75% (SD = 22.53), which was significantly below chance performance [$t(22) = -5.80$, $p < 0.0001$]. Mean percent long-VOT responses for the J-LONG training was 65.85% (SD = 24.30), which was significantly above change performance [$t(24) = 3.26$, $p = 0.003$]. Critically, percent long-VOT responses were higher for the J-LONG compared to the J-SHORT training group [$t(46) = 6.36$, $p < 0.001$; $d = 1.84$]. This pattern indicates that listeners learned Joanne's characteristic VOTs during the training phases.

b. Category goodness. In order to examine whether learning a talker's characteristic productions influences internal category structure, we examined performance during the goodness rating test phases. For each participant, mean goodness as /k/ rating was calculated for each token presented during the goodness test by collapsing across the six test phases. Figure 1(a) shows the mean goodness function for each training group derived by averaging across the participants within each group. Consider first the function for

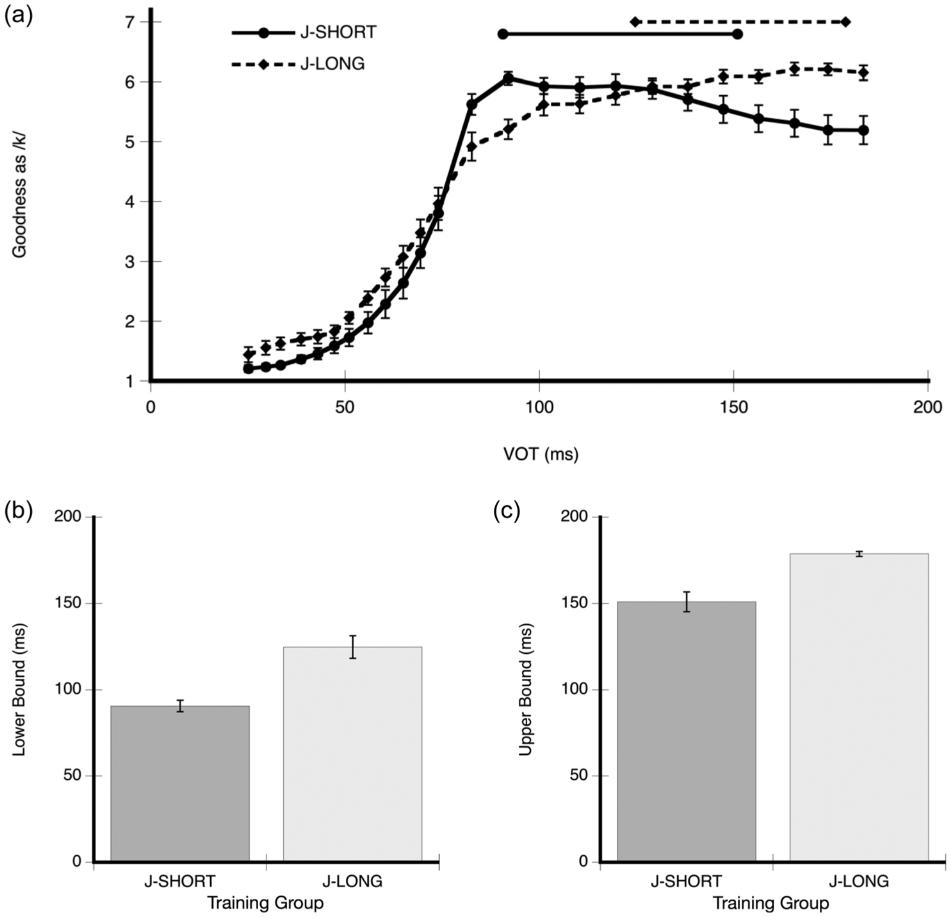

FIG. 1. Mean performance for the category goodness test phases for the J-SHORT and J-LONG training groups. (a) Shows the mean goodness functions and best exemplar ranges (indicated by the horizontal lines), (b) shows mean lower bound of the best exemplar region, and (c) shows the mean upper bound of the best exemplar region. Error bars indicate standard error of the mean.

the J-SHORT training group. Tokens with short VOTs were given extremely low goodness ratings, presumably because these VOTs were perceived as /g/. As VOTs increased so too did mean goodness ratings; however, only a small range of VOTs were given the highest ratings. Now consider performance for the J-LONG training group. As with the J-SHORT training, the goodness function shows a stable pattern such that tokens with short VOTs are given low goodness as /k/ ratings, ratings systematically increase as does VOT, but only a small range of VOTs are given the highest ratings. Critically, visual inspection of the two functions reveals that the range of VOTs that received the highest ratings are displaced between the two training groups, with the range of VOTs rated highest for the J-LONG training group located at longer VOTs compared to the range of VOTs rated highest for the J-SHORT training group.

To quantify this difference, we used standard convention to calculate a best exemplar region for each participant following previously outlined procedures (e.g., Allen and Miller, 2001; Volaitis and Miller, 1992). This procedure worked as follows. First, for each participant, we identified the peak goodness rating of his or her goodness function. This was used to define the best exemplar region, which was quantified as the range of VOTs corresponding to 90% of the peak. For example, if a participant had a mean rating of 7.0 for any token, then the best exemplar range would be defined as VOTs corresponding to ratings of 6.3 and higher. Using these criteria, we located the lower bound of the best exemplar region by identifying the VOT corresponding to when the goodness ratings first reached the best exemplar criterion. We located the upper bound of the best exemplar region by identifying the VOT corresponding to when goodness ratings first fell below the best exemplar criterion. In the event that the exact best exemplar criterion was not assigned to a VOT, linear interpolation between the two adjacent points was used to determine the VOT that would have received that rating. We imposed the constraint that to be taken as the lower (or upper) bound of the best exemplar region, the goodness ratings had to meet the best exemplar criterion for two of three consecutive tokens in order to measure category goodness with greater stability. If goodness ratings did not fall to such a degree that the upper bound of the best exemplar region could not be calculated, then we took the longest VOT presented (183 ms) as the measure of the upper bound of the best exemplar region. This was the case for 6 participants in the J-SHORT training group and 15 participants in the J-LONG training group.

Figure 1(b) shows the mean lower bound of the best exemplar range for the two training groups, with the mean upper bound shown in Fig. 1(c). In both cases, the best exemplars are located at longer VOTs for the J-LONG compared to the J-SHORT training group. We used analysis of variance (ANOVA) to examine this difference statistically. First, the mean lower bound of the best exemplar region was submitted to ANOVA with the factors of training group (J-SHORT or J-LONG) and test order (goodness-identification or identification-goodness). The results of the ANOVA showed a main effect of training group [$F(1,44) = 18.94$, $p < 0.001$; $\eta^2 = 0.298$], with the lower bound of the best exemplar region located at longer VOTs for the J-LONG compared to the J-SHORT training group. There was no main effect of test order [$F(1,44) = 0.18$, $p = 0.677$; $\eta^2 = 0.003$], nor was there an interaction between training group and test order [$F(1,44) = 0.38$, $p = 0.542$; $\eta^2 = 0.006$]. With respect to the upper bound of the best exemplar region, ANOVA showed a main effect of training group [$F(1,44) = 21.34$, $p < 0.001$; $\eta^2 = 0.329$], with the upper bound located at longer VOTs for the J-LONG compared to the J-SHORT training group. Again, there was no main effect of test order [$F(1,44) = 0.02$, $p = 0.892$; $\eta^2 = 0.000$], nor was there an interaction between training group and test order [$F(1,44) = 0.15$, $p = 0.700$; $\eta^2 = 0.002$]. These results suggest that experience with Joanne's voice during training promoted a comprehensive reorganization of internal category space; specifically, listeners adjusted category goodness to be centered on Joanne's characteristic productions.

c. Category identification. Performance for the category identification test phases was measured in order to determine whether exposure to Joanne's characteristic VOTs promoted a change in category boundary, as has been shown for exposure to a talker's productions that are ambiguous between two categories (e.g., Kraljic and Samuel, 2007). For each participant, we calculated mean percent /k/ responses for each VOT presented during test by collapsing across the six category identification test phases. Figure 2 shows the mean identification functions for both training groups. In order to determine whether the boundary or the slope of the identification functions differed between the two training groups, we used probit analyses to fit an ogive to the identification function for each individual participant. In all cases, the ogive was an excellent fit to responses, using r as an indicant ($r > 0.98$ in all cases). The mean of the ogive was used as a metric of the category boundary, showing the VOT that corresponded to 50% /k/ responses. The slope of the ogive was used as a metric of how categorical the function was, with increased slopes indicating a less categorical function. As shown in Fig. 2, visual inspection suggests that there was no reliable difference in either the boundary or the slope between the two training groups.

Mean category boundary was submitted to ANOVA with the factors of training group and test order. The results of the ANOVA showed no main effect of training group [$F(1,44) = 2.67$, $p = 0.109$; $\eta^2 = 0.041$], with the boundary placed at the same VOT for both training groups. There was, however, a significant main effect of order [$F(1,44) = 8.97$, $p = 0.004$; $\eta^2 = 0.139$] and an interaction between order and training group [$F(1,44) = 8.93$, $p = 0.005$; $\eta^2 = 0.138$]. *Post hoc* pairwise comparisons were conducted in order to explicate the nature of the interaction. There was no difference in category boundary as a function of test order for those in the J-SHORT training group [$t(21) = 0.01$, $p = 0.996$]. Test order did influence performance in the J-LONG training group, with the category boundary slightly shorter (65 versus 77 ms) for those who completed the identification test followed by the goodness test compared to those who completed these tests in the opposite order [$t(23) = -3.85$, $p < 0.001$]. If it were the case that exposure to Joanne's voice in

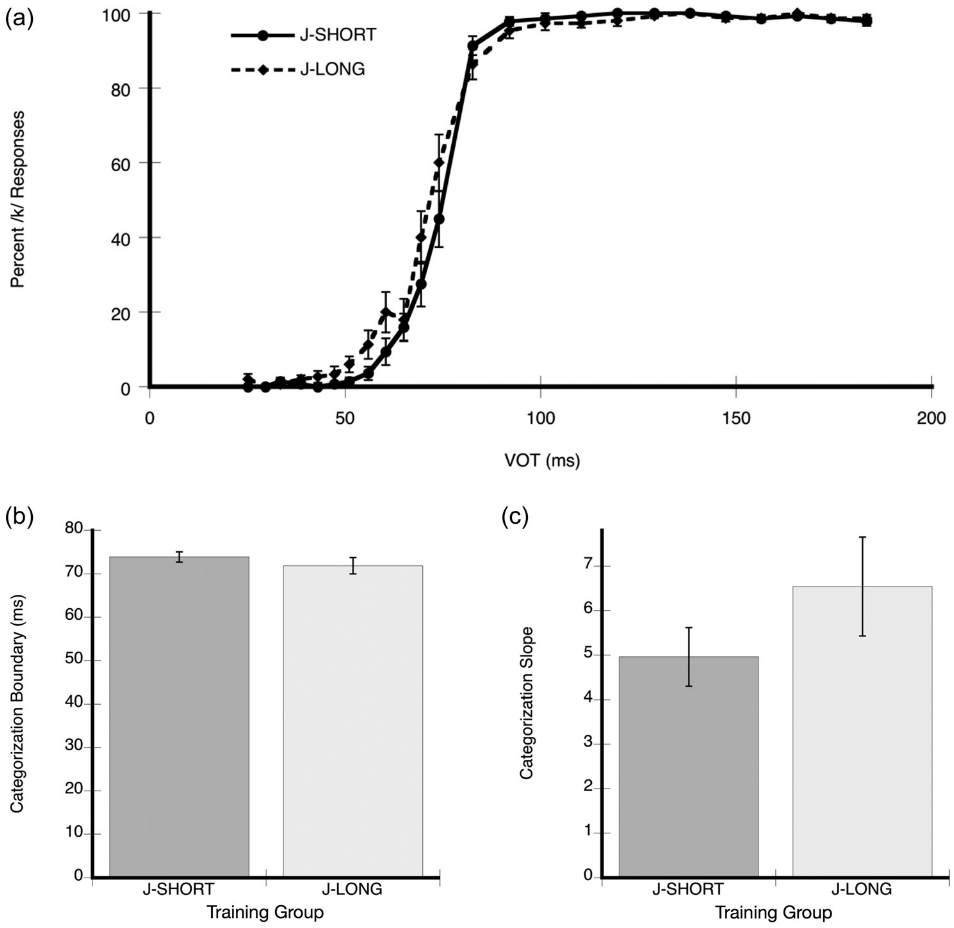

FIG. 2. Mean performance for the category identification test phases in Experiment 1 for the J-SHORT and J-LONG training groups. (a) Shows the mean identification functions, (b) shows mean category boundary, and (c) shows mean identification slope. Error bars indicate standard error of the mean.

the latter test order prior to completing the identification test influenced how listeners performed, then we would have expected to observe a similar effect in the J-SHORT training group. Because we do not, we interpret the order effect as spurious. (Consistent with this account, this was the only case for all analyses presented in this paper where we observed either a main effect of test order or an interaction with test order.)

A parallel ANOVA was performed on the identification slopes. The ANOVA showed no main effect of training group [$F(1,44) = 1.11, p = 0.297; \eta^2 = 0.024$], no main effect of order [$F(1,44) = 1.10, p = 0.302; \eta^2 = 0.024$], and no interaction between training group and order [$F(1,44) = 0.02, p = 0.878; \eta^2 = 0.001$]. Collectively, results from the category identification test phases provide no evidence that experience during training led to adjustments to the phonetic category boundary.

III. EXPERIMENT 2

The results of Experiment 1 indicated that listeners tracked talkers' characteristic VOTs, as predicted based on previous research (Allen and Miller, 2004; Theodore and Miller, 2010). Moreover, the results showed that listeners adjusted internal category structure to reflect the talker's characteristic production, while leaving the stop voicing category boundary intact. The goal of Experiment 2 was to examine whether this type of talker-specific perceptual learning generalizes to novel lexical items. Indeed, if—as suggested by results of Experiment 1—learning a talker's characteristic VOTs promotes a comprehension reorganization of the speech sound category, then the learning effect should not be limited to specific training tokens.

Two additional groups of listeners were tested using the procedures outlined for Experiment 1, with one exception. Though the training stimuli remained the same, the test stimuli consisted of novel lexical items. If perceptual learning for talker-specific phonetic detail as measured with respect to internal category structure generalizes to novel lexical items, as has been shown in other measures of talker-specific perceptual learning (Nygaard *et al.*, 1994; Theodore and Miller, 2010), then we predict that we will observe similar patterns for the generalization items tested in Experiment 2.

A. Methods

1. Participants

Thirty-four adults between the ages of 18 and 24 who did not participate in Experiment 1 were recruited following the previously outlined criteria. The sample size and stopping rule were determined based on the effect sizes observed in Experiment 1, which indicated that adequate statistical power could be achieved with fewer participants. Of the 34 participants, one was excluded due to failure to learn the talkers' voices, measured by talker identification accuracy less than 65% during the identification training sessions. Two additional participants were excluded because they were not monolingual. Of the remaining 31 participants, 15 were assigned to the J-SHORT training group and 16 were assigned to the J-LONG training group.

2. Stimuli

The training stimuli were identical to those used in Experiment 1. In order to assess generalization of learning during training, another continuum was created following the methods outlined previously. This continuum was also drawn from the stimuli reported in Theodore and Miller (2010). Specifically, a naturally-produced token of *goal* in Joanne's voice was used to generate a continuum that perceptually ranged from *goal* to *coal*. The naturally-produced token was selected to match VOT of the original *gain* token and was equated for word duration by trimming energy from the offset of the final consonant and was also equated for rms amplitude. Twenty-four tokens were selected from this continuum to serve as test stimuli by matching each token to the corresponding token used from the *gain–cane* continuum. The VOTs of the selected tokens ranged from 26 to 184 ms, with a step size of 4–5 ms for the first 12 tokens and a step size of 8–10 ms for the last 12 tokens. All 24 tokens were used in the goodness rating and category identification test phases. For the explicit memory test phases, we created pairs of stimuli as outlined previously, with each pair consisting of a short- and long-VOT variant of *coal* separated by 750 ms of silence. The VOTs of the short- and long-VOT variant were matched to those presented for the *cane* test pairs used in Experiment 1. Thus, training stimuli were identical to those used in Experiment 1 and consisted of the words *gain* and *cane* produced by Joanne and Sheila. All listeners were tested on Joanne's voice, as in Experiment 1, but were presented with the novel lexical items *goal* and *coal*, which matched the acoustic-phonetic characteristics to the test items used in Experiment 1.

3. Procedure

The procedure was identical to that used in Experiment 1, save that now listeners were trained using the *gain–cane* stimuli and tested using tokens from the *goal–coal* continuum.

B. Results

1. Training

As in Experiment 1, performance during training approached ceiling during the first six training phases. Mean accuracy for the phonetic decision was 96.70% (SD = 7.73) for Joanne's voice and 96.25% (SD = 4.27) for Sheila's voice. Mean accuracy for the talker decision was also high, 95.13% (SD = 6.07) and 93.62% (SD = 7.72) for Joanne and Sheila's voices, respectively.

2. Test

a. Explicit memory. Performance during test was analyzed separately for each of the three test types as outlined in Experiment 1. First, we examined performance during the explicit memory test phases by calculating, for each subject, percent long-VOT responses across the six test sessions. Mean percent long-VOT responses were 62.17% (SD = 23.52) for the J-LONG training group and 29.81% (SD = 23.21) for the J-SHORT training group, a difference that statistically reliable [$t(29) = 3.85$, $p < 0.001$; $d = 1.38$]. This pattern indicates that experience during training guided performance at test, even for the novel test item, such that listeners who heard Joanne produce /k/ with characteristically long VOTs during training chose more long VOT responses at test compared to listeners who heard Joanne produce /k/ with characteristically short VOTs.

b. Category goodness. Performance during the goodness rating test phases was analyzed as outlined in Experiment 1. The longest VOT presented (184 ms) was taken as the upper bound of the best exemplar region for four participants in the J-SHORT training group and 11 participants in the J-LONG training group. Figure 3 shows the mean goodness functions for both training groups [Fig. 3(a)], and the mean lower and upper bounds of the goodness functions [Figs. 3(b) and 3(c), respectively] across participants. As in Experiment 1, the range of VOTs rated most prototypical is located at longer VOTs for the J-LONG compared to the J-SHORT training group. To examine this difference statistically, the mean lower and upper bounds of the best exemplar range were examined in separate ANOVAs with the factors of training group and test order. For the lower bounds of the best exemplar range, results of ANOVA showed a main effect of training group, with the lower bound located at longer VOTs for the J-LONG compared to the J-SHORT training group [$F(1,27) = 5.47$, $p = 0.027$; $\eta^2 = 0.141$]. There was no reliable effect of test order [$F(1,27) = 3.41$, $p = 0.076$; $\eta^2 = 0.088$], nor an interaction between training group and test order [$F(1,27) = 2.98$, $p = 0.096$; $\eta^2 = 0.077$].

The results of the ANOVA for the upper bounds of the best exemplar range also showed a main effect of training group, with the upper bound located at longer VOTs for the J-LONG compared to the J-SHORT training group [$F(1,27) = 5.38$, $p = 0.028$; $\eta^2 = 0.160$]. The main effect of order [$F(1,27) = 0.01$, $p = 0.945$; $\eta^2 = 0.000$] and the interaction between training group and order [$F(1,27) = 1.24$, $p = 0.274$; $\eta^2 = 0.037$] were not reliable. These results suggest that the adjustments that listeners made to category goodness extend beyond the particular lexical item presented during training.

c. Category identification. As in Experiment 1, we examined whether exposure to Joanne's characteristic VOTs during training resulted in adjustments to the voicing boundary. To do so, we examined performance during identification test phases. Figure 4 shows the mean phoneme identification functions for both training groups [Fig. 4(a)], as well as the mean category boundary [Fig. 4(b)] and identification slopes [Fig. 4(c)] extracted for each participant as outlined for Experiment 1. With respect to the phoneme identification functions, visual inspection suggests that there was no difference between the training groups for either the category boundary or the identification slope. To test this,

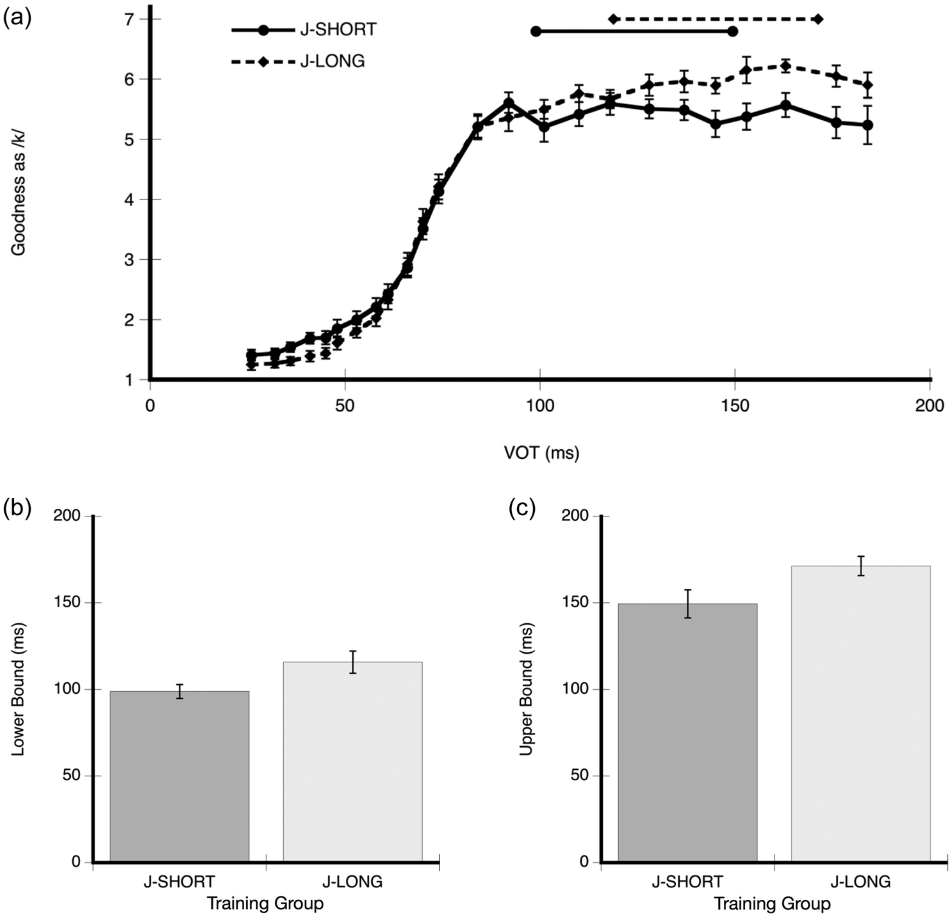

FIG. 3. Mean performance for the category goodness test phases for the J-SHORT and J-LONG training groups. (a) Shows the mean goodness functions and best exemplar ranges (indicated by the horizontal lines), (b) shows mean lower bound of the best exemplar region, and (c) shows the mean upper bound of the best exemplar region. Error bars indicate standard error of the mean.

mean boundaries and slopes were submitted to separate ANOVAs with the factors of training group and test order. With respect to the category boundary, there was no effect of training group [$F(1,27) = 0.02$, $p = 0.880$; $\eta^2 = 0.001$], order [$F(1,27) = 0.55$, $p = 0.465$; $\eta^2 = 0.020$], nor an interaction between the two factors [$F(1,27) = 0.32$, $p = 0.576$; $\eta^2 = 0.012$]. Parallel results were observed for the identification slopes; there was no main effect of training group [$F(1,27) = 0.00$, $p = 0.983$; $\eta^2 = 0.000$] or order [$F(1,27) = 1.70$, $p = 0.203$; $\eta^2 = 0.052$], and the interaction between the two was not reliable [$F(1,27) = 0.43$, $p = 0.518$; $\eta^2 = 0.015$]. These results mirror those observed in Experiment 1. Specifically, exposure to Joanne's characteristic VOTs, which represented clearly defined category members, promoted a reorganization of internal category structure but had no influence on the phonetic category boundary.

3. Transfer of learning

As described in Secs. II B and III B, performance for all test tasks was qualitatively similar when tested on the trained items (Experiment 1) and the novel lexical items (Experiment 2). One additional set of analyses was conducted that quantitatively compared performance between Experiment 1 and Experiment 2 in order to examine the magnitude of learning.

a. Explicit memory. Mean percent-long VOT responses was submitted to ANOVA with the factors of experiment (1 vs 2) and training group (J-SHORT vs J-LONG). As expected, the ANOVA showed a reliable main effect of training group [$F(1,75) = 48.78$, $p < 0.001$; $\eta^2 = 0.391$]. However, there was no main effect of experiment [$F(1,75) = 0.10$, $p = 0.755$; $\eta^2 = 0.001$], nor did experiment interact with training group [$F(1,27) = 0.99$, $p = 0.323$; $\eta^2 = 0.008$], indicating full transfer of learning in that performance for the novel lexical item was equivalent to the trained lexical item.

b. Category goodness. Two ANOVAs were performed with the factors of experiment and training group, one for the lower bound of the best exemplar region and one for the upper bound of the best exemplar region. Results for the two dependent measures were equivalent. Specifically, there was a main effect of training group on the location of the lower bound of the best exemplar region [$F(1,75) = 20.23$, $p < 0.001$; $\eta^2 = 0.207$], but no significant main effect of experiment [$F(1,75) = 0.00$, $p = 0.960$; $\eta^2 = 0.000$] or an interaction between training group and experiment [$F(1,75) = 2.28$, $p = 0.135$; $\eta^2 = 0.023$]. With respect to the location of the upper bound of the best exemplar region, there was a robust main effect of training group [$F(1,75) = 21.99$, $p < 0.001$; $\eta^2 = 0.224$]. No main effect of experiment was observed [$F(1,75) = 0.72$, $p = 0.398$; $\eta^2 = 0.007$], nor did experiment interact with training group [$F(1,75) = 0.30$, $p = 0.585$; $\eta^2 = 0.003$].

c. Category identification. Two ANOVAs were performed with the factors of experiment and training group,

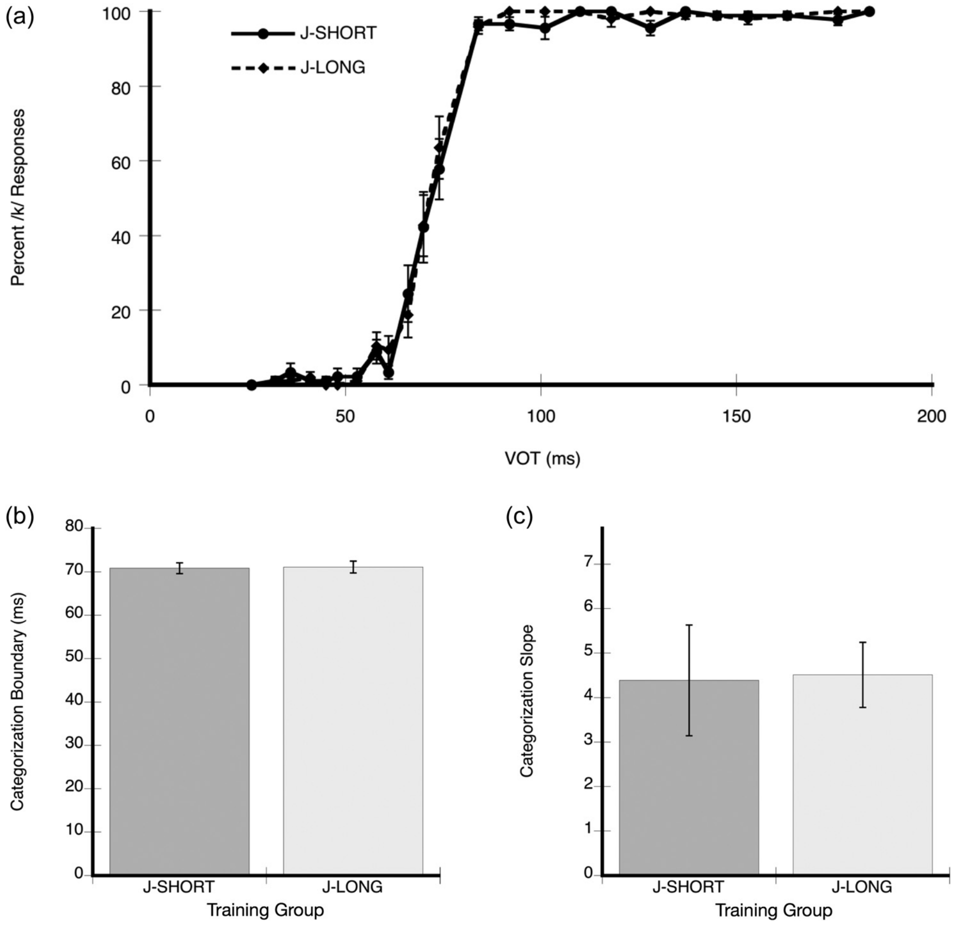

FIG. 4. Mean performance for the category identification test phases in Experiment 2 for the J-SHORT and J-LONG training groups. (a) Shows the mean identification functions, (b) shows mean category boundary, and (c) shows mean identification slope. Error bars indicate standard error of the mean.

one using the categorization boundary as the dependent measure and one using the categorization slope as the dependent measure. For the category boundary, we observed no main effect of training group [$F(1,75) = 0.31$, $p = 0.579$; $\eta^2 = 0.004$], no main effect of experiment [$F(1,75) = 1.30$, $p = 0.259$; $\eta^2 = 0.017$], nor an interaction between these two factors [$F(1,75) = 0.52$, $p = 0.474$; $\eta^2 = 0.007$]. The same pattern held for the categorization slope, with no main effect of training group [$F(1,75) = 0.72$, $p = 0.398$; $\eta^2 = 0.004$], no main effect of experiment [$F(1,75) = 1.69$, $p = 0.198$; $\eta^2 = 0.009$], nor an interaction between experiment and training group [$F(1,75) = 0.53$, $p = 0.470$; $\eta^2 = 0.003$].

Collectively, the results of the combined analysis suggest that for the metrics where talker-specific perceptual learning was observed in the individual experiments—tracking a talker's characteristic production and adjusting internal category structure to reflect that characteristic production–full transfer of learning occurred. These results confirm that this type of talker-specific perceptual learning promotes a comprehensive remapping from the acoustic-phonetic signal to speech sound category such that learning is not constrained to individual training items.

IV. DISCUSSION

A fundamental goal of research in the domain of speech perception is to describe how listeners reliably extract consonants and vowels from the acoustic stream given that there is no one-to-one relationship between the acoustic signal and an individual speech sound. A rich example of this lack of invariance concerns talker differences in phonetic properties of speech. Individual talkers have a unique phonetic signature, which listeners must accommodate in order to achieve stability in language comprehension. Research to date has highlighted mechanisms that promote such robust perception, including extensive evidence that listeners are able to dynamically adjust processing in light of systematic acoustic-phonetic variability (e.g., Nygaard *et al.*, 1994; Nygaard and Pisoni, 1998; Eisner and McQueen, 2005). Most notably, the literature on perceptual learning for speech has shown that listeners will use lexical information to adjust the phonetic boundaries between individual speech sounds in order to incorporate new members into an established phonetic category (e.g., Norris *et al.*, 2003).

In the current work, we examined another potential way that listeners may accommodate a talker's phonetic signature. Specifically, we examined whether listeners would modify internal category structure in light of a talker's characteristic productions, as has been shown for other systematic sources of acoustic-phonetic variation (e.g., Volaitis and Miller, 1992). The productions we presented to listeners were unambiguous, well-defined category members that contrast with the type of idiosyncratic talker productions examined in the lexically-informed perceptual learning literature. The results showed that listeners robustly shifted internal category structure to be centered on previous experience with the talker's voice. However, there was no evidence that experience during training modified the category boundary. These findings held even when the word presented during test differed from that presented during training, suggesting that perceptual reorganization generalized beyond the particular items presented during training. We do note, however, that the scope of learning may have been attenuated for the novel compared to the trained word, given that the effect sizes observed in Experiment 2 were smaller than those observed in Experiment 1. These results, when considered with respect to previous literature on perceptual learning for speech, suggest that the ways in which listeners adjust speech sound representations for individual talkers may in fact depend on the nature of the production to be accommodated. Specifically, the results collectively suggest that listeners may be conservative in shifting phonetic category boundaries, doing so only when they must resolve ambiguity in the signal. In contrast, listeners may accommodate unambiguous productions through a reorganization of perceptual space within the category proper, leaving the boundary intact.

The current results constrain theoretical models of spoken language processing by pointing toward distinct learning outcomes for talker-specific productions—and presumably other sources of acoustic-phonetic variability including novel dialects and foreign accents—depending on the nature of production to be incorporated into existing category space. Such input-driven constraints on functional plasticity for speech perception may help explain how listeners maintain stability of linguistic knowledge while simultaneously showing flexibility for phonetic representations. To complete this account, however, examination of the degree to which boundary adjustments are decoupled from modifications to internal category structure is warranted. Specifically, it is not yet known whether perceptual learning for ambiguous productions that leads to adjustments to category boundaries is limited to the boundary region or whether it also promotes reorganization within the category proper. In addition, future work needs to examine the degree to which lexical support influences talker-specific perceptual learning. In the current work, talkers' characteristic productions were always embedded in real English words. Accordingly, lexical access could occur. Research on lexically-informed perceptual learning has shown that the category boundary adjustments listeners make to accommodate a talker's ambiguous productions do no occur when the productions are embedded in nonwords (Norris *et al.*, 2003). A mechanistic account of perceptual organization for talker-specific phonetic variability will be promoted by considering whether the effects observed in the current work hold if talkers' characteristic, unambiguous productions are embedded in nonwords. Moreover, the current work generates predictions for the neural basis of talker-specificity effects in speech perception. Previous work has suggested a division of labor among frontal and temporal regions for phonetic category processing, with the former processing phonetic ambiguity for tokens near the phonetic category boundary, and the latter processing internal phonetic category structure (Myers, 2007). These same two neural structures show sensitivity to talker-specific variants, when that sensitivity is altered via exposure to boundary-value tokens embedded in unambiguous lexical contexts (Myers and Mesite, 2014). The current finding that the internal structure of phonetic categories, but not the category boundary, is altered by exposure to unambiguous tokens suggests that superior temporal regions, but not frontal regions, will respond to talker-specific changes in internal category structure. Future work is aimed at addressing these questions.

ACKNOWLEDGMENTS

Portions of this manuscript were completed as a Master's thesis by J.A.L. under the direction of R.M.T. This research was funded by a Faculty Large Grant (University of Connecticut Research Foundation) and by faculty start-up funds (University of Connecticut, College of Liberal Arts and Sciences) to R.M.T.

Allen, J. S., and Miller, J. L. (**2001**). "Contextual influences on the internal structure of phonetic categories: A distinction between lexical status and speaking rate," Percept. Psychophys. **63**, 798–810.

Allen, J. S., and Miller, J. L. (**2004**). "Listener sensitivity to individual talker differences in voice-onset-time," J. Acoust. Soc. Am. **115**, 3171–3183.

Cooper, F. S., Delattre, P. C., Liberman, A. M., Borst, J. M., and Gerstman, L. J. (**1952**). "Some experiments on the perception of synthetic speech sounds," J. Acoust. Soc. Am. **24**, 597–606.

Eisner, F., and McQueen, J. (**2005**). "The specificity of perceptual learning in speech processing," Percept. Psychophys. **67**, 224–238.

Fant, G. (**1993**). "Some problems in voice source analysis," Speech Comm. **13**, 7–22.

Klatt, D. H., and Klatt, L. C. (**1990**). "Analysis, synthesis, and perception of voice quality variations among female and male talkers," J. Acoust. Soc. Am. **87**, 820–857.

Kraljic, T., and Samuel, A. G. (**2005**). "Perceptual learning for speech: Is there a return to normal?," Cogn. Psychol. **51**, 141–178.

Kraljic, T., and Samuel, A. G. (**2007**). "Perceptual adjustments to multiple speakers," J. Mem. Lang. **56**, 1–15.

Kraljic, T., Samuel, A. G., and Brennan, S. E. (**2008**). "First impressions and last resorts: How listeners adjust to speaker variability," Psychol. Sci. **19**, 332–338.

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

Miller, J. (**1994**). "On the internal structure of phonetic categories: A progress report," Cognition **50**, 271–285.

Miller, J. L., Grosjean, F., and Lomanto, C. (**1984**). "Articulation rate and its variability in spontaneous speech: A reanalysis and some implications," Phonetica **41**, 215–225.

Murray, I. R., and Arnott, J. L. (**1993**). "Toward the simulations of emotion in synthetic speech," J. Acoust. Soc. Am. **93**, 1097–1108.

Myers, E. B. (**2007**). "Dissociable effects of phonetic competition and category typicality in a phonetic categorization task: An fMRI investigation," Neuropsychologia **45**, 1463–1473.

Myers, E. B., and Mesite, L. M. (**2014**). "Neural systems underlying perceptual adjustments to non-standard speech tokens," J. Mem. Lang. **76**, 80–93.

Newman, R. S., Clouse, S. A., and Burnham, J. L. (**2001**). "The perceptual consequences of within-talker variability in fricative production," J. Acoust. Soc. Am. **109**, 1181–1196.

Norris, D., McQueen, J. M., and Cutler, A. (**2003**). "Perceptual learning in speech," Cogn. Psychol. **47**, 204–238.

Nygaard, L. C., and Pisoni, D. B. (**1998**). "Talker-specific learning in speech perception," Percept. Psychophys. **60**, 355–376.

Nygaard, L. C., Sommers, M. S., and Pisoni, D. B. (**1994**). "Speech perception as a talker-contingent process," Psychol. Sci. **5**, 42–46.

Peterson, G. E., and Barney, H. L. (**1952**). "Control methods used in a study of the vowels," J. Acoust. Soc. Am. **24**, 175–184.

Sumner, M. (**2011**). "The role of variation in the perception of accented speech," Cognition **119**, 131–136.

Theodore, R. M., and Miller, J. L. (**2010**). "Characteristics of listener sensitivity to talker-specific phonetic detail," J. Acoust. Soc. Am. **128**, 2090–2099.

Theodore, R. M., Miller, J. L., and DeSteno, D. (**2009**). "Individual talker differences in voice-onset-time: Contextual influences," J. Acoust. Soc. Am. **125**, 3974–3982.

Volaitis, L. E., and Miller, J. L. (**1992**). "Phonetic prototypes: Influence of place of articulation and speaking rate on the internal structure of voicing categories," J. Acoust. Soc. Am. **92**, 723–735.