Characteristics of listener sensitivity to talker-specific phonetic detail[a)]

Rachel M. Theodore[b)] and Joanne L. Miller
Department of Psychology, Northeastern University, Boston, Massachusetts 02115

(Received 2 December 2009; revised 24 June 2010; accepted 4 July 2010)

Previous research shows that listeners are sensitive to talker differences in phonetic properties of speech, including voice-onset-time (VOT) in word-initial voiceless stop consonants, and that learning how a talker produces one voiceless stop transfers to another word with the same voiceless stop [Allen, J. S., and Miller, J. L. (2004). J. Acoust. Soc. Am. **115**, 3171–3183]. The present experiments examined whether transfer extends to words that begin with different voiceless stops. During training, listeners heard two talkers produce a given voiceless-initial word (e.g., *pain*). VOTs were manipulated such that one talker produced the voiceless stop with relatively short VOTs and the other with relatively long VOTs. At test, listeners heard a short- and long-VOT variant of the same word (e.g., *pain*) or a word beginning with a different voiceless stop (e.g., *cane* or *coal)*, and were asked to select which of the two VOT variants was most representative of a given talker. In all conditions, which variant was selected at test was in line with listeners' exposure during training, and the effect was equally strong for the novel word and the training word. These findings suggest that accommodating talker-specific phonetic detail does not require exposure to each individual phonetic segment. © *2010 Acoustical Society of America.* [DOI: 10.1121/1.3467771]

PACS number(s): 43.71.Bp, 43.71.Es [AJ] Pages: 2090–2099

I. INTRODUCTION

A major goal of research in the domain of speech perception has been to describe how listeners extract stable linguistic percepts given that the acoustic-phonetic information produced for individual speech segments, and thus for individual words, varies considerably from utterance to utterance. Factors contributing to systematic variability in the speech signal are numerous and include surrounding phonetic context (Delattre *et al.*, 1955), speaking rate (Miller, 1981), and even idiosyncratic pronunciation differences among talkers (e.g., Allen *et al.*, 2003; Hillenbrand *et al.*, 1995; Klatt, 1986; Newman *et al.*, 2001; Peterson and Barney, 1952). Listeners' accommodation of the latter source of variability, talker-specific phonetic detail, is the focus of the current experiments.

It is now known that listeners retain in memory many surface characteristics of the speech signal (Church and Schacter, 1994; Nygaard *et al.*, 2000; Palmeri *et al.*, 1993; Schacter and Church, 1992), including the phonetic signature associated with individual talkers' voices (e.g., Goldinger, 1998; see also Remez *et al.*, 1997). Findings from the domain of spoken word recognition have shown that talker-specific phonetic variability can be used to customize speech processing for individual talkers. Talker familiarity has been shown to increase intelligibility (Bradlow and Bent, 2008; Nygaard *et al.*, 1994) and decrease processing time (Clarke and Garrett, 2004). These effects hold when listeners learn to identify talkers on the basis of isolated words (Nygaard *et al.*, 1994) or sentences (Nygaard and Pisoni, 1998), and can be achieved even with short periods of exposure (Bradlow and Pisoni, 1999; Clarke and Garrett, 2004).

These findings provide evidence that listeners use talker-specific phonetic detail to facilitate word recognition. Although relatively little is known about which aspects of the speech signal listeners encode at the level of individual talkers, and how such encoding subsequently facilitates word recognition, there is some evidence suggesting that the talker-specificity effects observed at higher levels of processing may reflect, at least in part, adjustments that listeners make at a prelexical, or segmental, level of representation.

For example, Norris *et al.* (2003) proposed one way in which listeners might perceptually adjust for at least some talker differences in speech production. The type of idiosyncratic production they examined was ambiguous production of individual speech sounds that may be found in, for example, foreign-accented speech. In their experiments, listeners were exposed to an ambiguous fricative midway between /f/ and /s/ during a lexical decision training phase. For some listeners, the ambiguous fricative was presented in the context of /f/-final words, such that perceiving it as /f/ supported lexical recognition but perceiving it as /s/ did not. For other listeners, the ambiguous fricative was presented in the context of /s/-final words, such that perceiving it as /s/ supported lexical recognition but perceiving it as /f/ did not. At test, all listeners were asked to categorize members of an /f/-/s/ continuum. The results showed that listeners adjusted phonetic boundaries so as to include ambiguous tokens within the

[a)] Portions of this work were presented at Acoustics '08, a joint meeting of the Acoustical Society of America, the European Acoustics Association, and the Société Française d'Acoustique (Paris, France, July 2008), and at the 156th Meeting of the Acoustical Society of America (Miami, Florida, November 2008).

[b)] Author to whom correspondence should be addressed. Current address: Department of Cognitive and Linguistic Sciences, Brown University, Box 1978, Providence, Rhode Island 02912. Electronic mail: rachel_theodore@brown.edu phonetic category that supported lexical recognition. Subsequent work has shown that accommodation of idiosyncratic fricative productions extends to novel words that contain the same fricative (McQueen *et al.*, 2006). Moreover, subsequent work has also shown that the lexically-informed boundary adjustment is sometimes applied on a talker-specific basis (Eisner and McQueen, 2005; Kraljic and Samuel, 2005; but see Kraljic and Samuel, 2007) and results from minimal exposure to a talker's productions (Kraljic and Samuel, 2006). These findings raise the possibility that some of the adjustments listeners make in order to accommodate talker-specific phonetic variation occur at a prelexical level of representation.[1]

Lexically-informed perceptual learning is one process that may underlie rapid adjustment to differences in production across talkers, particularly when adjusting to talkers whose pronunciations are so deviant that they fall near a category boundary and could be perceived as more than one speech sound. Yet, many of the acoustic-phonetic differences found across talkers involve well-defined category members, rather than members near a category boundary (Allen *et al.*, 2003; Newman *et al.*, 2001; Peterson and Barney, 1952). Talkers can produce different acoustic instantiations that are unambiguously identified as the same speech sound and it is likely that listeners encounter these differences more often than the ambiguous productions that may be found in, for example, foreign-accented speech.

One central issue concerns whether or not listeners can accommodate such fine-grained differences in production across talkers; that is, when the particular segment in question is unambiguous and well within a phonetic category. This issue has recently been addressed for the phonetically relevant property of voice-onset-time (VOT), which is an articulatory property of stop consonants that is measured acoustically as the time between the onset of the release burst of the stop consonant and the onset of periodicity associated with subsequent vocal fold vibration (Lisker and Abramson, 1964). In English, VOT is an important marker of the voicing contrast, distinguishing voiced /b d g/ from their voiceless counterparts /p t k/. Of particular relevance to the current work is the finding that individual talkers differ in their characteristic VOTs for voiceless stops; controlling for speaking rate, some talkers produce longer VOTs than other talkers (Allen *et al.*, 2003). Moreover, recent research indicates that these talker differences are stable across a change in place of articulation (Theodore *et al.*, 2009).

Allen and Miller (2004) examined whether listeners can track such talker differences in VOT, focusing on the alveolar voiceless stop /t/. In their experiments, listeners participated in training and test phases. In the training phases, listeners learned to identify the voices of two talkers, "Annie" and "Laura." On a single trial during the training phases, listeners were presented with the word *dime* or *time* and were asked to identify the voice of the talker and the initial phoneme of the word. Critically, the VOTs of the *time* tokens were manipulated. While both VOTs clearly specified the initial /t/, one talker had relatively short VOTs and the other had relatively long VOTs. On a single trial during the test phases, listeners were presented with two variants of *time* produced by one of the talkers, a short-VOT variant and a long-VOT variant, and were asked to identify which of the two variants was more typical of that talker. Results showed that which token listeners chose at test depended on their previous exposure to that talker's voice. For example, if they had heard Annie produce short VOTs during training, they chose the short-VOT variant of Annie's speech at test. Likewise, if they heard Annie produce long VOTs during training, they chose the long-VOT variant of Annie's speech at test. Moreover, the effect persisted when listeners were tested on the novel word *town*. Transfer to a novel word was replicated in an additional experiment in which listeners were exposed to *town* during the training phases, and then tested on *time*.

That listeners transferred information learned about a talker's characteristic VOTs to a novel word indicates that talker-specific VOT was tracked in some way that was not dependent on a particular training stimulus. This finding suggests that exposure to one lexical item can potentially inform the listener as to how that talker produces many other lexical items, at least those items that begin with the same stop consonant. The issue addressed in the current research is whether the scope of generalization extends beyond a given stop consonant. In particular, we asked whether listeners who learn how a particular talker produces /p/ also learns how that talker produces /k/. As noted above, talker differences in VOT are stable across place of articulation. Specifically, talkers who produce /p/ with relatively long (or short) VOTs also produce /k/ with relative long (or short) VOTs (Theodore *et al.*, 2009). Thus, if listeners can transfer information learned in the context of one voiceless stop to another, they would be informed as to that talker's characteristic productions for a much larger set of lexical items than if no such cross-segment transfer occurred.

We examined the issue of cross-segment transfer in two experiments that used a slightly modified version of the Allen and Miller (2004) paradigm. Experiment 1 examined transfer in a minimal pair context and Experiment 2 examined transfer in a non-minimal pair context. In each experiment, listeners participated in two sessions. Within each session, listeners alternated between training and test phases. The test phases in Session 1 examined performance for the word presented during training, and the test phases in Session 2 examined performance for a novel word that began with a different voiceless stop than was presented during training. The critical question was whether listeners would demonstrate sensitivity to talkers' characteristic VOTs for the novel word tested in Session 2, and, if so, to what degree. In the extreme, if listeners can fully transfer information learned in the context of one voiceless stop to a different voiceless stop, then test performance during Session 2 should be at the same level as test performance during Session 1.

II. EXPERIMENT 1

In Experiment 1, we provided the simplest test of transfer; namely, we examined transfer between words that form minimal pairs. During training, listeners heard two female talkers, "Annie" and "Laura," produce *pain*. Speech synthesis techniques were used to differentially manipulate Annie and Laura's characteristic VOTs such that one group of listeners heard Annie produce relatively short VOTs and Laura produce relatively long VOTs, and the other group of listeners heard Annie produce relatively long VOTs and Laura produce relatively short VOTs.

All listeners participated in two sessions of alternating training and test phases. Training phases across the two sessions were the same, and involved listeners being exposed to appropriate VOT variants of *pain* for each talker, and the voiced-initial counterpart *bane*. Critically, the test phases of the two sessions differed. In Session 1, listeners were tested on *pain* and in Session 2 they were tested on a novel voiceless-initial word, *cane*. On each trial at test, listeners were presented with a short-VOT and long-VOT variant of the test word produced by one of the talkers and were asked to select which variant was most representative of that particular talker.

Based on Allen and Miller (2004), for Session 1 of each experiment we expected that which VOT variant was selected at test for the word presented during training would be contingent on previous exposure to the talkers' characteristic VOTs. The critical questions, tested in Session 2, were: (1) Would exposure during training influence which VOT variant was selected for the novel word at test? (2) If so, would it be to the same degree as that observed for the training word?

A. Method

1. Subjects

Twenty subjects were recruited for participation in the experiment. Half of the subjects were assigned to the A-SHORT/L-LONG training group and the other half were assigned to the A-LONG/L-SHORT training group. All subjects were native speakers of English between the ages of 18 and 45, with no reported speech or hearing disorders. Subjects were either paid or received partial course credit for their participation. Any subject who did not correctly identify the two talkers' voices during training or who did not correctly identify the voiced-initial and voiceless-initial tokens presented during training was replaced with a new subject, as described in the results section.

2. Stimulus preparation

The stimuli consisted of two sets of tokens, a labial-initial *bane*/*pain* set and a velar-initial *gain*/*cane* set. Each set contained synthesized versions of the voiced-initial and voiceless-initial words that were based on the speech of two female talkers. Within each set, multiple variants of the voiceless-initial word were created such that they differed from one another in VOT. Stimulus preparation was based on the procedure outlined in Allen and Miller (2004), and the reader is referred to that paper for comprehensive details on the preparation procedure. We give and overview of the procedure here.

Many female talkers produced 20 repetitions of the words *bane*, *gain*, and *goal* (recorded for use in Experiment 2), along with many fillers. Their speech was recorded via microphone (AKG C460B) onto digital audiotape in a sound-attenuated booth. All recordings were digitized at a sampling rate of 20 KHz using the CSL system (KayPENTAX). A waveform of each repetition of *bane* and *gain* was generated with the Praat speech analysis software (Boersma, 2001); using this display, VOT and word duration were measured to the nearest millisecond. VOT was measured from the release burst to the onset of high-amplitude, periodic energy associated with the vowel, and word duration was measured from the release burst to the offset of periodic energy associated with the final consonant.

TABLE I. VOT values (ms) of the *bane*/*pain* training stimuli.

Training group: A-short/L-long			
		pain	
Talker	*bane*	Token 1	Token 2
Annie	0	60	69
Laura	0	155	164
Training group: A-long/L-short			
		pain	
Talker	*bane*	Token 1	Token 2
Annie	0	155	165
Laura	0	60	68

Two talkers (different from those used in Allen and Miller, 2004) were selected; the talkers are referred to as Annie and Laura. The selected talkers had roughly comparable overall word durations, and, as confirmed by analyses presented in the results section, the two talkers had perceptually distinct voices. One repetition of *bane* and one repetition of *gain* were selected from each talker such that VOT for a given word was approximately matched across the talkers. The four selected tokens were equated for word duration by deleting from the final consonant such that all tokens were 568 ms in duration, and were then equated for root-mean-square (RMS) amplitude.

The ASL system (KayPENTAX) was used to perform a pitch-synchronous LPC analysis on each of the four selected tokens. The output of this analysis was used to create a synthesized version of each selected token, and, using the synthesized *bane* and *gain* tokens from each of the two talkers, four VOT series were created by systematically changing parameters of the LPC analysis and synthesizing new tokens using the modified parameters. This procedure yielded, for each talker, one series of stimuli that perceptually ranged from *bane* to *pain* and one series that ranged from *gain* to *cane*, thus creating a pool of tokens that were matched on overall duration and differed in word-initial VOT.

Five tokens were selected from each *bane*/*pain* VOT series to serve as training stimuli, including one voiced-initial token and four voiceless-initial tokens. The VOT values of the selected tokens are shown in Table I for each training group. The particular tokens were selected to include one voiced token (the first step of each series), two tokens from the short-VOT voiceless region that were two steps apart on the continuum (to simulate naturally occurring within-talker variability), and, likewise, two tokens from the long-VOT voiceless region that were two steps apart on the

TABLE II. VOT values (ms) of the *pain*, *cane*, and *coal* test stimuli. The pain and cane test stimuli were used in Experiment 1 and the pain and coal test stimuli were used in Experiment 2.

	pain		*cane*		*coal*	
Talker	Short-VOT	Long-VOT	Short-VOT	Long-VOT	Short-VOT	Long-VOT
Annie	65	160	88	181	87	185
Laura	64	160	88	183	88	182

continuum. The particular short-VOT and long-VOT tokens were selected in order to maximize the difference in VOT between these tokens, while ensuring that VOTs of the short-VOT tokens were not so short that they fell within the ambiguous VOT region of a particular continuum and VOTs of the long-VOT tokens were not so long so as to yield extreme exemplars of the particular voiceless stop. Moreover, the VOTs of the selected short-VOT and long-VOT tokens were closely matched across talkers.

Both training groups were presented with the same test stimuli. Two tokens were selected from each *bane*/*pain* VOT series for use during test in Session 1, including one short-VOT voiceless-initial token and one long-VOT voiceless-initial token. The VOT values of the selected tokens are shown in Table II. The particular tokens were selected based on the training stimuli. Recall that the two short-VOT voiceless tokens and the two long-VOT voiceless tokens from each series selected for training were two steps apart on the continuum. The intermediate token in all cases was selected for use during test. In addition, two *cane* tokens were selected from each *gain*/*cane* VOT series for use during test in Session 2. These tokens were selected such that the difference in VOT between the short-VOT and long-VOT variants was approximately the same as that of the *pain* test tokens, and that the selected tokens adhered to the well known influence of place of articulation on VOT in speech production, with VOTs for labial stops being shorter than those for velar stops (e.g., Cho and Ladefoged, 1999; Lisker and Abramson, 1964).

As described in Allen and Miller (2004), the synthesis techniques used to generate the VOT series result in a potential amplitude-based confound in that tokens with shorter VOTs have higher overall amplitude (measured in terms of RMS level) than tokens with longer VOTs. This potential confound was eliminated by generating two amplitude variants (high and low) for each selected token and presenting both amplitude variants during training and test in order to ensure that subjects' performance could not be attributed to the amplitude difference of the short-VOT and long-VOT tokens. At presentation, amplitude of the high and low variants was 67 dB SPL and 65 dB SPL, respectively.

For the *bane*/*pain* stimulus set, separate training lists were created for the A-SHORT/L-LONG training group and the A-LONG/L-SHORT training group, using both amplitude variants of each training stimulus. The training lists for the A-SHORT/L-LONG training group contained Annie and Laura's *bane* tokens, Annie's short-VOT *pain* tokens, and Laura's long-VOT *pain* tokens. The training lists for the A-LONG/L-SHORT training group contained, in analogous fashion, Annie and Laura's *bane* tokens, Annie's long-VOT *pain* tokens, and Laura's short-VOT *pain* tokens. In each training list, an extra *bane* token was included so as to equate the number of voiced-initial and voiceless-initial tokens within the list. Thus, a training list consisted of 16 tokens (2 talkers X 2 *bane* tokens X 2 *pain* tokens X 2 amplitude levels) in randomized order. Sixteen such lists were created for presentation to the listeners during training across the two sessions of the experiment.

Separate test lists were created for the *bane*/*pain* and *gain*/*cane* stimulus sets. For each stimulus set, separate test lists were created for Annie and Laura, with each test list consisting of pairs of each talker's test stimuli. Each pair consisted of the appropriate short-VOT and long-VOT test stimulus, separated by 750 ms of silence. Each stimulus was presented at two amplitude levels, with the amplitude level on a given trial held constant, and the order of the short-VOT and long-VOT variants counterbalanced across trials. This resulted in four pairings of test stimuli for each talker. A test list consisted of a randomized sequence of two repetitions of these pairings, resulting in eight trials for each test list. In total, eight test lists were created for each stimulus set, four for each talker. The *pain* test lists were used in Session 1 and the *cane* test lists were used in Session 2.

3. Procedure: Session 1

As noted earlier, 20 subjects participated in Experiment 1, with half assigned to the A-SHORT/L-LONG training group and half assigned to the A-LONG/L-SHORT training group. Testing took place in a sound-attenuated booth, with auditory stimuli presented via headphones (Sony MDR-V6). All subjects alternated between training and test phases. During training, subjects were presented with the *bane*/*pain* training lists according to their training group. At test, subjects were presented with the *pain* test lists. The overall session consisted of three main components: familiarization, practice, and the experiment proper.

a. Familiarization. Subjects first completed a brief familiarization component involving the stimuli to be used in the experiment proper. The main purpose of this component was for listeners to learn to identify each talker's voice. One familiarization list (16 tokens), composed in the same manner as the training lists created for the experiment proper, was presented. Each trial consisted of the auditory presentation of the stimulus followed by visual presentation of the name of the talker who produced that stimulus. The name of the talker appeared on a computer display 750 ms after the offset of the auditory stimulus, and remained on the screen for 1500 ms. The next trial began following a pause of 2000 ms. Subjects were instructed to listen to each word and view the name of the talker in order to learn to identify each talker's voice. Subjects did not provide any responses during familiarization.

b. Practice. After familiarization, subjects completed a brief practice component in order to be exposed to the training and test tasks. The practice component was blocked by talker, with the order of the talkers counter-balanced within each training group. For each talker block, subjects completed practice training and practice test. For the practice training, they were presented with one list (24 tokens) consisting of three randomized blocks of the eight stimuli to be used during training in the experiment proper for that talker (2 voiced-initial tokens X 2 voiceless-initial tokens X 2 amplitude variants). On each trial, they were asked to identify the initial consonant, indicating their response by pressing a button labeled B or P on a response keypad. No feedback was provided. For the practice test within each talker block, subjects were presented with one test list (8 trials) composed in the same manner as the test lists created for the experiment proper. Subjects were instructed to indicate which of the two VOT variants presented on each trial was most representative of that talker's voice. They indicated their response by pressing a button labeled 1 if they thought it was the first member of the pair and a button labeled 2 if they thought it was the second member of the pair. No feedback was provided during test.

c. Experiment proper. Following familiarization and practice, the experiment proper began with the alternation between training phases and test phases that used the lists described above in the Stimulus Preparation section. In each training phase, subjects were presented with one training list (16 trials) with the order of the training lists determined randomly for each subject. Subjects were asked to identify, for each stimulus, both the talker and the initial consonant. They indicated their response by pressing one of four buttons labeled Annie B, Annie P, Laura B, and Laura P. Feedback was provided for the talker choice only, in the form of a visual display that showed YES for a correct response and NO and the name of the talker for an incorrect response. The visual feedback appeared 750 ms after the button response and remained on the screen for 1500 ms. The next trial began following a pause of 2000 ms.

During test, subjects were presented with one of the test lists for one of the talkers (8 trials). The order of presentation for the test lists was determined randomly for each subject, with the constraint that no more than three tests lists of the same talker were presented in a row. Instructions during the test phase were the same as those described above for the practice test phase. At the beginning of each test phase, the talker's name for that test phase appeared on the screen. On each test trial, subjects indicated which of the two VOT variants presented on the trial was most representative of that talker's voice. The pause between trials was 2000 ms, timed from the button response.

Overall, the sequence within the session was as follows: familiarization, practice training/practice test for one talker, practice training/practice test for the other talker, training phase, test phase, and additional alternation between training and test phases to the completion of eight test phases. Subjects were given a short break after the completion of four test phases.[2]

4. Procedure: Session 2

Following a brief break after Session 1, subjects completed Session 2, which was the transfer session. In Session 2, subjects completed an additional eight alternations of training phases and test phases. Training stimuli remained the same as presented in Session 1 (*bane* and *pain*), but listeners were tested on the novel *cane* test lists. The procedural details for the training and test phases followed those outlined for Session 1.

B. Results

Training. Performance during training was analyzed separately for talker and phonetic identification by calculating percent correct responses. For talker identification, a response was considered correct if the talker was identified, even if the initial consonant was not. For phonetic identification, a response was considered correct if the initial consonant was identified, even if the talker was not. Mean percent correct for both talker and phonetic identification was calculated for each subject, for each session. High performance during training was necessary for inclusion in the study. A criterion of 80% correct in each session for both talker identification and phonetic identification was adopted to indicate high performance. Two subjects were replaced because they did not reach the criterion for talker identification. For the 20 subjects included in the experiment, performance during training across both sessions was near ceiling for both talker identification (95%) and phonetic identification (99%).

Test. The data were analyzed following the procedure outlined in Allen and Miller (2004). Performance during test was analyzed in terms of percent long-VOT responses. [Recall that on each trial during test, listeners selected either the short-VOT or the long-VOT variant of *pain* (Session 1) or *cane* (Session 2); because percent short-VOT and long-VOT responses must sum to 100, quantifying performance in terms of both is redundant.] For each subject, mean percent long-VOT responses was calculated for a given talker separately for each session. Figure 1 shows percent long-VOT responses for each talker separately for each training group, with Session 1 responses shown in the top panel and Session 2 responses shown in the bottom panel. With our measure of performance, percent long-VOT responses, sensitivity to talker differences in VOT will manifest as an interaction between training group and talker. As described in detail below, two sets of analyses were performed for each session. In the primary analysis, ANOVA was used to examine the statistical significance of the critical interaction between training group and talker. In the secondary analysis, a set of planned comparisons was used to confirm that the interaction was due to the predicted pattern of results.[3]

Consider first performance during Session 1. The percentage of long-VOT responses was submitted to ANOVA with the between-subjects factor of training group and the

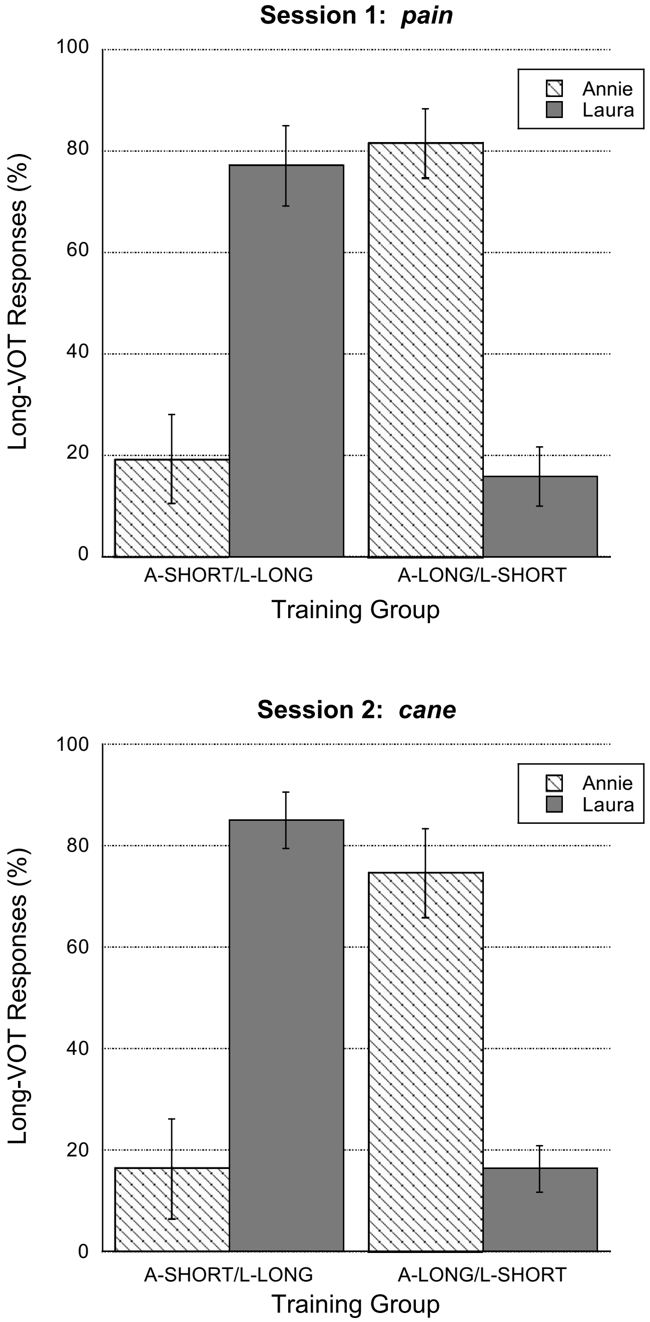

FIG. 1. Mean percent long-VOT responses for the test phases in Experiment 1 for each training group, for each talker's voice. Session 1 data are shown in the top panel and session 2 data are shown in the bottom panel. Error bars indicate standard error of the mean.

within-subjects factor of talker. Results of the ANOVA confirmed a significant interaction between training group and talker $[F(1,18)=42.06, p<.001]$. As stated above, a secondary set of analyses was conducted in order to confirm the nature of the interaction. First, we wanted to confirm that percentage of long-VOT responses for each talker within each training group was different from chance, which was 50% in each case. For these one-sample tests, we used the t distribution with $df=9$, $\alpha=0.05$. Results of these tests did confirm that performance was different from chance. Second, planned comparisons were performed to ensure that the interaction revealed in the ANOVA was due to our predicted pattern of results. The planned comparisons consisted of comparing performance for Annie and Laura's voice within each training group (within-subjects), as well as comparing performance for each talker's voice across the two training groups (between-subjects). For the within-subjects comparisons, we used the t distribution with $df=9$, $\alpha=0.05$. For the between-subjects comparisons, we used the t distribution with $df=18$, $\alpha=0.05$. The results from the planned comparisons confirmed that the interaction was due to the predicted pattern of results. Specifically, there were fewer long-VOT responses for Annie's voice compared to Laura's voice in the A-SHORT/L-LONG training group, and this pattern was reversed for listeners in the A-LONG/L-SHORT training group. Additionally, there were fewer long-VOT responses for Annie's voice from listeners in the A-SHORT/L-LONG training group compared to listeners in the A-LONG/L-SHORT training group, and this pattern was reversed for Laura's voice. This pattern of results, as predicted, confirms that which VOT variant was selected at test in Session 1 was contingent on exposure to a talker's characteristic VOTs during training.

To address the central question as to whether or not tracking a talker's VOTs transfers across place of articulation, we examined percent long-VOT responses selected during Session 2, shown in the bottom panel of Fig. 1. As shown in the figure, performance in Session 2 was similar to performance in Session 1, indicating that transfer across place of articulation did occur. The results from the primary analysis showed a significant interaction between training group and talker $[F(1,18)=46.27, p<.001]$, and the results from the secondary set of analyses confirmed that the percentage of long-VOT responses for each talker within each training group was different from chance, and that the interaction was due to the predicted pattern of performance. These results indicate that even for the novel word *cane*, listeners used experience with the talker's voices provided in the context of *pain* to guide which VOT variant was selected at test.

In order to assess the strength of the transfer, an additional ANOVA was performed using the factors of training group, talker, and session (within-subjects). The ANOVA confirmed a significant interaction between talker and training group $[F(1,18)=52.02, p<.001]$, as expected. The effect of session was not significant $[F(1,18)<1]$, and, critically, there were no significant interactions with session (in all cases, $p>.10$). These results indicate full transfer of learning between the voiceless stops in that performance at test was as robust for the novel word as it was for the training word.

In Experiment 1, we provided the simplest test of transfer across place of articulation; namely, the only phonological difference between training and test words was the initial stop. In Experiment 2, we examined whether cross-segment transfer is limited to this constrained environment or, instead, whether transfer – and even complete transfer – would also be observed between words that are phonologically less similar.

III. EXPERIMENT 2

The design of Experiment 2 was analogous to that of Experiment 1. The only difference was that we increased the phonological distance between training and novel test words such that they no longer formed minimal pairs, potentially increasing the difficulty of transfer of talkers' characteristic VOTs across place of articulation. Listeners were trained on *bane/pain* in both Sessions 1 and 2, and they were tested on *pain* in Session 1 and on *coal* in Session 2. As in Experiment 1, two questions were considered: Would exposure during training influence which VOT variant was selected for the novel word at test? If so, would it be to the same degree as that observed for the training word?

A. Method

1. Subjects

Twenty different subjects were recruited for participation in the experiment following criteria outlined for Experiment 1. Half of the subjects were assigned to the A-SHORT/L-LONG training group and the other half were assigned to the A-LONG/L-SHORT training group.

2. Stimulus preparation

The stimuli consisted of two sets of tokens, including the labial-initial *bane/pain* set used in Experiment 1 and an additional velar-initial *goal/coal* set. Preparation of the *goal/coal* set followed the procedures outlined for Experiment 1, as summarized below.

Stimuli. One token of *goal* was selected for each talker from the recordings described for Experiment 1. Both tokens were trimmed to 568 ms in duration in order to equate word duration to the *bane/pain* stimulus set. Synthesized versions of the *goal* tokens were made using the ASL system, and using these synthesized tokens, a VOT series ranging from *goal* to *coal* was generated for each talker.

Two tokens were selected from each series to be presented during test in Session 2, a short-VOT *coal* token and a long-VOT *coal* token. The VOTs of the selected test tokens are shown in Table II. As in Experiment 1, a high- and low-amplitude variant were generated for each token in order to eliminate a potential amplitude-based confound. At presentation, amplitude of the high and low variants for the *goal/coal* set was 71 dB SPL and 69 dB SPL, respectively. (Due to intrinsic vowel differences, RMS amplitude of the high and low *coal* variants was increased by 4 dB relative to the amplitude of the bane/pain tokens in order to match loudness across the two sets of stimuli.) Following the procedures used in Experiment 1, test lists for the experiment were constructed using the selected *coal* stimuli.

3. Procedure

The procedural details for Experiment 2 were the same as those outlined for Experiment 1; only the stimuli changed. During training, subjects were presented with the *bane/pain* training lists according to their training group. At test, subjects in Experiment 2 were presented with the *pain* test lists in Session 1 and the *coal* test lists in Session 2.

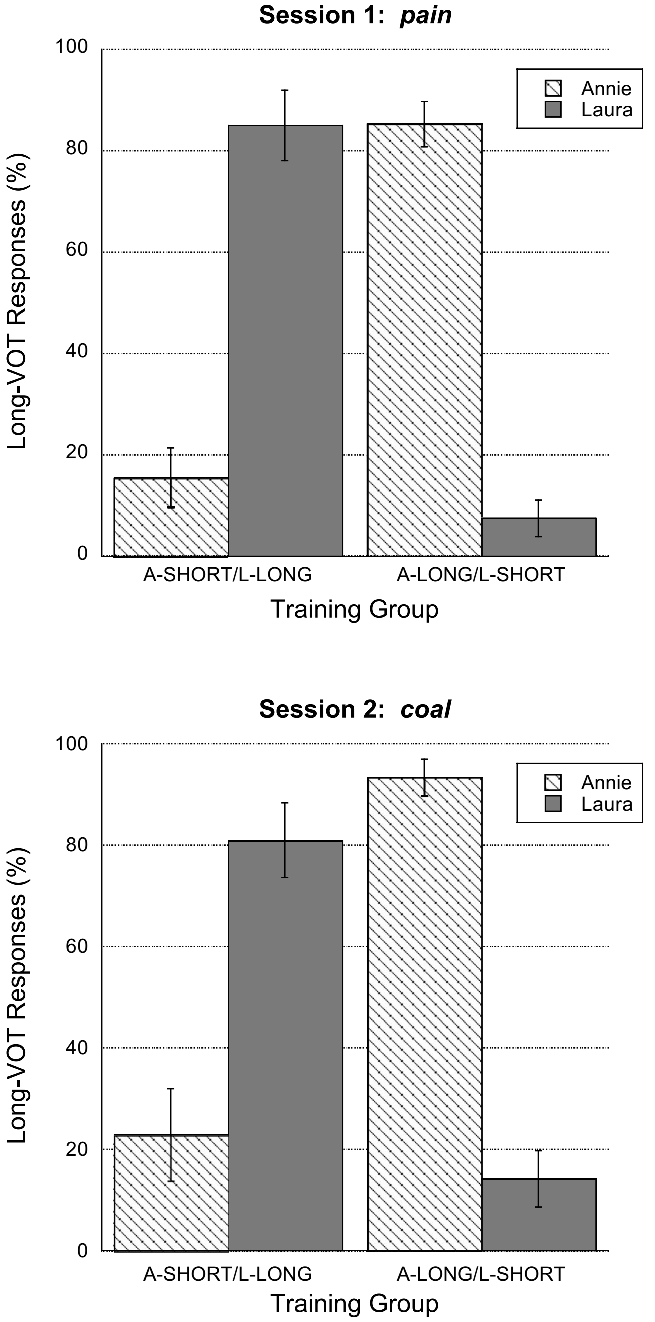

FIG. 2. Mean percent long-VOT responses for the test phases in Experiment 2 for each training group, for each talker's voice. Session 1 data are shown in the top panel and session 2 data are shown in the bottom panel. Error bars indicate standard error of the mean.

B. Results

Training. Performance during training was analyzed as outlined for Experiment 1. Two subjects were replaced for sub-criterion performance on talker identification. For the 20 subjects included in the experiment, performance during training was near ceiling for both talker identification (96%) and phonetic identification (99%).

Test. Fig. 2 shows percent long-VOT responses for each talker separately for each training group, with Session 1 responses shown in the top panel and Session 2 responses shown in the bottom panel. The pattern of performance seen here is the same as was observed in Experiment 1. Specifically, listeners selected the VOT variant at test that was in line with previous exposure to the talkers' voices for both the word presented during training and the novel word. To confirm the statistical significance of this pattern, the primary and secondary analyses outlined in Experiment 1 were performed on the percentage of long-VOT responses for each session. Results of the primary analysis for Session 1 and for Session 2 confirmed the predicted interaction between talker and training group [$F(1,18)=118.59, p<.001$; $F(1,18)=58.49, p<.001$; respectively]. Results from the secondary analyses confirmed that in both sessions the percentage of long-VOT responses for each talker in each training group was different from chance, and that the interactions were due to the predicted pattern of results.

As in Experiment 1, one additional ANOVA was performed in order to assess the strength of the transfer using the factors of training group, talker, and session. The expected interaction between talker and training group was confirmed [$F(1,18)=96.63, p<.001$]. The effect of session approached significance [$F(1,18)=3.39, p<.10$], but, critically, there were no significant interactions with session ($p>.10$ in all cases). This pattern of results indicates full transfer of learning between the training word and the novel word.[4]

IV. DISCUSSION

The acoustic signal of speech is highly variable. As reviewed in the Introduction, there is much evidence within the domain of speech perception indicating that listeners accommodate for this variability, at least in part, by retaining in memory fine-grained information regarding the acoustic instantiation of individual speech segments and using this information to facilitate speech processing (e.g., Goldinger, 1996). One source of information used by the perceptual system is idiosyncratic differences in speech production associated with individual talkers. This has been demonstrated for higher levels of processing, including word recognition (e.g., Nygaard *et al.*, 1994), as well as for lower levels of processing, including segmental perception (e.g., Norris *et al.*, 2003).

In terms of segmental perception, recent findings indicate that listeners can track talker differences in phonetic properties of speech. Focusing on VOT in word-initial voiceless stop consonants, Allen and Miller (2004) showed that for a given voiceless stop, listeners could learn that one talker produced characteristically short VOTs and that a different talker produced characteristically long VOTs. This finding provided the basis for the current work, which examined the scope of generalization underlying such sensitivity to talker differences in VOT. Two experiments were conducted. In both experiments, two groups of listeners were differentially exposed to characteristic VOTs for two talkers; one talker produced short VOTs and the other talker produced longer VOTs. Exposure was provided during training phases in which listeners heard both talkers produce a voiceless stop consonant in the context of a word. Sensitivity to talkers' characteristic VOTs was assessed for the word presented during training and for a novel word that began with a different voiceless stop than that presented during training. Across the two experiments, we manipulated the phonological distance between the training and novel words: the words formed minimal pairs (*pain* and *cane*) in Experiment 1 but formed non-minimal pairs (*pain* and *coal*) in Experiment 2.

The same pattern of results was found for both experiments. Specifically, sensitivity to talkers' characteristic VOTs was observed not only for the word presented during training, but also for the novel word. Moreover, for both the minimal pair and non-minimal pair cases, complete transfer of learning was obtained in that the magnitude of listener sensitivity to characteristic VOTs when tested on the novel word was equal to that observed when tested on the training word. These findings indicate that listeners do not require exposure to each individual segment in order to tune into a talker's phonetic signature; rather, there is generalization across similar segments. One striking aspect of the talker-specificity effects at lexical levels of processing, described in the Introduction, is that the processing advantage achieved by talker familiarity also generalizes to novel items (Nygaard and Pisoni, 1998). Such broad scope of generalization, at both the segmental and lexical levels, potentially affords more efficient accommodation of talker-specific phonetic detail compared to a learning process that operates in a segment-by-segment fashion, and may in fact underlie other findings indicating that adaptation to this type of variability in the speech signal is a rapid process (Clarke and Garrett, 2004).

The current demonstration of full transfer of talker-specific information about VOT across a change in place of articulation is consistent with at least two different learning mechanisms underlying such transfer. One possibility is that coding a talker's characteristic VOTs is linked to a phonetic feature. In this case, what listeners learned is how the two talkers implemented the feature voiceless for one stop consonant, and they were able to apply this knowledge to a voiceless stop produced at a different place of articulation. A strict feature-based account would predict full transfer of information, and that is what we found. Another possibility is that acoustic similarity underlies the transfer observed in the current work. On this account, listeners may have, for example, selected the novel variant of the voiceless stop that most closely matched the duration of the low amplitude, aperiodic energy associated with the VOT of the voiceless stop presented during training. On such an account, the magnitude of transfer would presumably vary in accord with the degree of similarity along the appropriate acoustic dimension between the training and test segment. The current finding of full transfer is consistent with such an account, given the assumption that the training and test segments were sufficiently similar to produce the same magnitude of effect, at least within the limits of the measure we used. Future research, which systematically varies acoustic similarity between training and test items, could help adjudicate the issue.

Another issue for future research concerns how listeners might track characteristic VOTs of individual talkers across changes in speaking rate. In the experiments reported here, as well as in Allen and Miller (2004), speaking rate (specified as word duration) was held constant. It is well known, however, that talkers frequently alter their rates of speech (Miller *et al.*, 1984), and that VOTs produced for word-initial voiceless stop consonants systematically increase as speaking rate slows (e.g., Miller *et al.*, 1986). Interestingly, recent findings have shown that precisely how much VOT increases as rate slows varies across individual talkers, with some talkers showing a more extreme increase than others (Theodore *et al.*, 2009). As a consequence, there are cases in which a talker who produces short VOTs relative to another talker at one speaking rate produces long VOTs relative to that talker at a different speaking rate. This is quite different from the case of place of articulation where, as noted earlier, talkers who produce a given voiceless stop with relatively long (or short) VOTs also produce other voiceless stops with relative long (or short) VOTs (Theodore *et al.*, 2009). Thus, unlike the minimal exposure required to promote transfer of information about a talker's characteristic VOTs across place of articulation, listeners potentially require exposure to a range of speaking rates in order to transfer such knowledge to a novel speaking rate. Just how listeners accomplish this remains to be determined.

Finally, the current findings, together with those of Allen and Miller (2004), provide clear evidence that listeners can track talker differences in phonetic properties of speech even when the talker-specific variation falls well within a phonetic category, such that there is no ambiguity in category membership (or identity of the lexical item). A question that remains is whether this ability reflects only learning about particular characteristic stimuli within the category, or instead involves a more comprehensive change in the mapping between acoustic signal and phonetic category, affecting all stimuli within a given category. It has long been known that phonetic categories are internally structured, with some category members perceived to be better exemplars of the category than others (e.g., Kuhl, 1991; Miller and Volaitis, 1989; Samuel, 1982). Of particular relevance to the current work, the internal structure of phonetic categories has been shown to systematically change in accord with how various acoustic-phonetic contextual factors alter the speech signal in production (e.g., Allen and Miller, 2001). It may be that the internal structure of phonetic categories is also tuned in accord with characteristic productions of individual talkers, so that as listeners become familiar with a talker's voice, they fine-tune phonetic categories on a talker-specific basis. Future research, which directly measures how talker familiarity affects perceived goodness of a range of within-category members, should help answer this question.

ACKNOWLEDGMENTS

This research was supported by NIH Grant R01 DC000130 to JLM and by NIH Grant F31 DC009114 (Ruth L. Kirschstein NRSA for Individual Predoctoral Fellows) to RMT, and formed the basis for part of the first author's doctoral dissertation at Northeastern University under the direction of the second author. We thank committee members Neal J. Pearlmutter and David DeSteno for their advice throughout the course of this project, and Janelle LaMarche for assistance with stimulus creation and data collection.

[1] Although Allen and Miller (2004) provided evidence that listeners are sensitive to talker differences in VOT, Kraljic and Samuel (2006, 2007) failed to observe talker-specificity in terms of listeners' accommodation of a novel stop voicing contrast that was implemented, in part, by VOT. This discrepancy may be explained by one of the many differences between the two paradigms that include using explicit versus implicit memory tasks, whether or not speaking rate was held constant, the amount of exposure provided to listeners, and whether VOT was manipulated independently of other aspects of the signal. A more theoretically interesting difference between the two paradigms concerns the nature of the productions presented to listeners; specifically, Allen and Miller examined sensitivity to well-defined exemplars of a given phonetic category whereas Kraljic and Samuel examined listeners' ability to incorporate an ambiguous exemplar into a phonetic category. Future research is needed to specify the conditions in which sensitivity to talker differences in VOT will be observed, as well as the conditions in which it may not be observed. Though Kraljic and Samuel failed to observe talker-specificity in terms of listeners' accommodation of a novel stop voicing contrast, they did show generalization across place of articulation. As described in the main text, the current experiments examine whether such transfer will also be observed in cases of talker-specific processing.

[2] Subjects alternated between training and test phases in order to minimize the effects of test on talker-specific memory. Recall that at test, listeners were exposed not only to VOTs for a particular talker that were in line with their experience during training, but also to VOTs that differed from their experience during training. Inasmuch as this additional exposure becomes part of listeners' memory for the talker, it is possible that long test phases could have altered their overall memory of a particular talker's VOT. Under this account, performance during the test phase would reflect not only exposure from training, but also exposure from test. In order to minimize this possibility, listeners alternated between longer training phases and shorter test phases.

[3] As described in the main text, the critical prediction tested in all experiments manifests in an ANOVA as a significant interaction between training group and talker. We made no predictions regarding the main effects of these factors and, indeed, in all cases the main effects of training group and talker were not significant ($p > .10$).

[4] In addition to the experiments presented in the main text, two additional experiments were conducted in order to ensure that transfer of learning between voiceless stop consonants is not contingent on the particular direction of transfer; that is, we wanted to ensure that listeners would also transfer from velar /k/ to labial /p/. To this end, two experiments were conducted following the methodology outlined in the main text. In one experiment, listeners were presented with VOT variants of *cane* during training, and were tested on *cane* in Session 1 and on *pain* in Session 2. In the other experiment, listeners were presented with VOT variants of *coal* during training, and were tested on *coal* in Session 1 on *pain* in Session 2. The results of both experiments were the same, and parallel to the findings presented in the main text. Specifically, the predicted interaction between training group and talker was observed in each session, and the magnitude of the interaction was the same across the two sessions for each experiment. These results replicate full transfer of learning between voiceless stop consonants, and confirm that this effect is not contingent on a particular direction of transfer.

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.

Allen, J. S., Miller, J. L., and DeSteno, D. (**2003**). "Individual talker differences in voice-onset-time," J. Acoust. Soc. Am. **113**, 544–552.

Boersma, P. (**2001**). "Praat, a system for doing phonetics by computer," Glot Int. **5**, 341–345.

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

Bradlow, A. R., and Pisoni, D. B. (**1999**). "Recognition of spoken words by native and non-native listeners: Talker-, listener-, and item-related factors," J. Acoust. Soc. Am. **106**, 2074–2085.

Cho, T., and Ladefoged, P. (**1999**). "Variations and universals in VOT: Evidence from 18 languages," J. Phonetics **27**, 207–229.

Church, B. A., and Schacter, D. L. (**1994**). "Perceptual specificity of auditory priming: Implicit memory for voice intonation and fundamental frequency," J. Exp. Psychol. Learn. Mem. Cogn. **20**, 521–533.

Clarke, C. M., and Garrett, M. F. (**2004**). “Rapid adaptation to foreign-accented English,” J. Acoust. Soc. Am. **116**, 3647–3658.

Delattre, P. C., Liberman, A. M., and Cooper, F. S. (**1955**). “Acoustic loci and transitional cues for consonants,” J. Acoust. Soc. Am. **27**, 769–773.

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

Goldinger, S. D. (**1996**). “Words and voices: Episodic traces in spoken word identification and recognition memory,” J. Exp. Psychol. Learn. Mem. Cogn. **22**, 1166–1183.

Goldinger, S. D. (**1998**). “Echoes of echoes? An episodic theory of lexical access,” Psychol. Rev. **105**, 251–279.

Hillenbrand, J., Getty, L. A., Clark, M. J., and Wheeler, K. (**1995**). “Acoustic characteristics of American English vowels,” J. Acoust. Soc. Am. **97**, 3099–3111.

Klatt, D. H. (**1986**). “The problem of variability in speech recognition and in models of speech perception,” in *Invariance and variability in speech processes*, edited by J. S. Perkell and D. H. Klatt (Erlbaum, Hillsdale, NJ), pp. 300–319.

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

Kraljic, T., and Samuel, A. G. (**2006**). “Generalization in perceptual learning for speech,” Psychon. Bull. Rev. **13**, 262–268.

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

Kuhl, P. K. (**1991**). “Human adults and human infants show a ‘perceptual magnet effect’ for the prototypes of speech categories, monkeys do not,” Percept. Psychophys. **50**, 93–107.

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

McQueen, J. M., Cutler, A., and Norris, D. (**2006**). “Phonological abstraction in the mental lexicon,” Cogn. Sci. **30**, 1113–1126.

Miller, J. L. (**1981**). “Effects of speaking rate on segmental distinctions,” in *Perspectives on the Study of Speech*, edited by P. D. Eimas and J. L. Miller (Erlbaum, Hillsdale, NJ), pp. 39–74.

Miller, J. L., Green, K. P., and Reeves, A. (**1986**). “Speaking rate and segments: A look at the relation between speech production and speech perception for the voicing contrast,” Phonetica **43**, 106–115.

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.

Miller, J. L., and Volaitis, L. E. (**1989**). “Effect of speaking rate on the perceptual structure of a phonetic category,” Percept. Psychophys. **46**, 505–512.

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., Burt, S. A., and Queen, J. S. (**2000**). “Surface form typicality and asymmetric transfer in episodic memory for spoken words,” J. Exp. Psychol. Learn. Mem. Cogn. **26**, 1228–1244.

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.

Palmeri, T. J., Goldinger, S. D., and Pisoni, D. B. (**1993**). “Episodic encoding of voice attributes and recognition memory for spoken words,” J. Exp. Psychol. Learn. Mem. Cogn. **19**, 309–328.

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

Remez, R. E., Fellowes, J. M., and Rubin, P. E. (**1997**). “Talker identification based on phonetic information,” J. Exp. Psychol. Hum. Percept. Perform. **23**, 651–666.

Samuel, A. G. (**1982**). “Phonetic prototypes,” Percept. Psychophys. **31**, 307–314.

Schacter, D. L., and Church, B. A. (**1992**). “Auditory priming: Implicit and explicit memory for words and voices,” J. Exp. Psychol. Learn. Mem. Cogn. **18**, 915–930.

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.