January 01, 2009

Some characteristics of talker-specific phonetic detail

Rachel Marie Theodore
Northeastern University

SOME CHARACTERISTICS OF TALKER-SPECIFIC PHONETIC DETAIL

A dissertation presented

by

Rachel M. Theodore

to

The Department of Psychology

In partial fulfillment of the requirements for the degree of
Doctor of Philosophy

in the field of

Psychology

Northeastern University
Boston, Massachusetts
February, 2009

SOME CHARACTERISTICS OF TALKER-SPECIFIC PHONETIC DETAIL

by

Rachel M. Theodore

ABSTRACT OF DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Psychology
in the Graduate School of Arts and Sciences of
Northeastern University, February, 2009

ABSTRACT

Talkers differ in the acoustic-phonetic information used to convey individual consonants and vowels. For many years, talker differences in phonetic properties of speech were considered as problematic noise for the perceptual system. Indeed, traditional accounts of speech perception posit that talker-specific phonetic detail is removed from the signal in the process of accessing abstract linguistic representations (e.g., Studdert-Kennedy, 1976). These accounts are challenged, however, by findings from multiple research domains indicating that talker-specific phonetic detail is retained in memory and can be used to facilitate speech processing (e.g., Goldinger, 1996; Nygaard & Pisoni, 1998). In order to develop a theoretical account of speech perception that describes how listeners accommodate talker differences in phonetic properties of speech, additional data on talker-specific phonetic detail are necessary from both the perception and production domains.

This dissertation examines talker-specific phonetic detail for the case of voice-onset-time (VOT) in word-initial voiceless stops. Previous research has shown that, when holding other influences on VOT constant, talkers differ in their characteristic VOTs with some talkers having shorter VOTs than other talkers (Allen et al., 2003). Other research has shown that listeners are sensitive to such talker differences in VOT in that they can learn, for a given voiceless stop, that one talker produces short VOTs and a different talker produces longer

VOTs (Allen & Miller, 2004). This dissertation consists of two projects that extend these findings.

The first project examined the scope of generalization underlying listener sensitivity to talker differences in word-initial 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 one voiceless stop consonant, either /p/ or /k/, in the context of a word (e.g., *pain* or *cane*). In test phases, listeners were presented with a short-VOT and a long-VOT variant of the word presented during training as well as a novel word that began with a different voiceless stop than presented during training. In both cases, listeners were asked to select which of the two VOT variants was most representative of a given talker. Across the two experiments, the phonological distance between the training and novel words was manipulated; the words formed minimal pairs (*pain* and *cane*) in Experiment 1 and non-minimal pairs (*pain* and *coal*) in Experiment 2. The same pattern of results was found in both experiments. Specifically, listeners selected the VOT variant in line with exposure during training not only for the word presented during training, replicating earlier findings (Allen & Miller, 2004), but also for the novel word. Moreover, for both the minimal pair and non-minimal pair cases, the magnitude of listener sensitivity to characteristic VOTs was the same for the novel word and the training word. These findings indicate that learning a talker's characteristic VOTs does not necessitate exposure to each phonetic segment; rather, there is transfer across similar segments.

In order to better inform theoretical accounts of the types of exposure listeners may require to transfer talker-specific phonetic detail across various dimensions, additional data from the production domain are necessary. To this end, the second project examined talker-specific phonetic detail in speech production. As stated above, talkers differ in VOT in word-initial stop consonants (Allen et al., 2003). Previous research also indicates that VOT is robustly affected by contextual influences, including speaking rate and place of articulation (e.g., Lisker & Abramson, 1964; Kessinger & Blumstein, 1997). This project examined whether these contextual influences on VOT are themselves talker-specific. Across two experiments, many tokens of labial /p/, alveolar /t/, and velar /k/ were elicited from talkers across a range of rates. All tokens formed syllables consisting of the voiceless stop followed by the vowel /i/ (e.g, /pi/). VOT and vowel duration (a metric of rate) were measured for each token. Hierarchical linear modeling analyses showed that: (1) VOT increased as rate slowed for all talkers, as expected, but the magnitude of the increase varied significantly across talkers; thus the effect of rate on VOT was talker-specific; (2) the talker-specific effect of rate was stable across a change in place of articulation; and (3) for all talkers VOTs were shorter for labial than for velar stops, as expected, and there was no significant variability in the magnitude of this displacement across talkers; thus the effect of place on VOT was not talker-specific. These findings provide basic information on how two contextual factors influence VOT at a talker-specific level and, in so doing, point to constraints on how listeners might accommodate such contextual variation when customizing phonetic categories for an individual talker's speech.

ACKNOWLEDGMENTS

Foremost, I wish to acknowledge Dr. Joanne L. Miller, my advisor. By way of example, she has taught me many lessons that will benefit my future endeavors. In particular, her commitment to careful attention to experimental detail and her dedication to conveying scientific findings with extreme clarity have taught me a valuable approach to conducting research. In addition, her ability to maintain a rigorous research program, be a top-notch educator, and be active within the broader academic community will provide both a standard and a model for me throughout my career. Dr. Miller's enthusiasm for her work is a joy to be near, and I have had lots of fun making science with her.

Gratitude is also extended to Dr. Neal J. Pearlmutter and Dr. David DeSteno, members of my dissertation committee, for their valuable input throughout the development of this dissertation.

As a member of the Speech Perception Lab, I have been fortunate to work with an amazing group of people. I acknowledge fellow graduate students Michele Mondini and Sandra Schwab for their camaraderie during our times of overlap in the Lab. Gratitude is also extended to Janelle LaMarche and Eliza Floyd, fabulous research technicians whose assistance in stimulus creation, subject running, and acoustic analysis has been most appreciated. Thanks is also given to the many outstanding undergraduate research assistants who have helped in many facets of the research process; notably, I acknowledge Katrina Smith for her assistance with the acoustical measurements for the experiments presented in Chapter 2. Furthermore, thanks is given to Joe Heck, whose Lab Wizardry skills have been instrumental in making sure that the equipment and programs necessary for collecting data were operational.

Finally, gratitude is extended to the Psychology Department at Northeastern University. The outstanding faculty and graduate students have created a supportive and stimulating environment for education. In particular, gratitude is extended to Dr. Rhea T. Eskew for providing such rigorous instruction in statistical hypothesis testing, as well as for taking an interest in my career after class had passed. Also, I acknowledge David Richters, Robert Griffo, and Nadya Vasilyeva for years of fascinating conversation regarding philosophy, science, and the philosophy of science, as well as for their friendship.

CONTENTS

INTRODUCTION

Talkers differ in the acoustic-phonetic information used to convey individual consonants and vowels. For many years, talker differences in phonetic properties of speech were considered as problematic noise for the perceptual system. Indeed, traditional accounts of speech perception posit that talker-specific phonetic detail is removed from the signal in the process of accessing abstract linguistic representations (e.g., Studdert-Kennedy, 1976). These accounts are challenged, however, by findings from multiple research domains indicating that talker-specific phonetic detail is retained in memory and can be used to facilitate speech processing (e.g., Goldinger, 1996; Nygaard & Pisoni, 1998). In order to develop a theoretical account of speech perception that describes how listeners accommodate talker differences in phonetic properties of speech, additional data on talker-specific phonetic detail are necessary from both the perception and production domains.

This dissertation examined talker-specific phonetic detail, focusing on voice-onset-time (VOT) in word-initial voiceless stops. The experiments in Chapter 1 examine the scope of generalization underlying listener sensitivity to talker differences in phonetic properties of speech. The experiments in Chapter 2 further characterize the nature of talker-specific phonetic detail in the acoustic signal of speech. The research presented in the two chapters is complementary; however, each chapter is written as a separate manuscript for publication and, as such, each is intended to stand on its own.

Chapter 1

Listener sensitivity to talker-specific phonetic detail

1.1 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 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 & Barney, 1952). Given the myriad acoustic-phonetic information that can be produced for a given speech segment, the task for the listener is essentially one of categorization, wherein many physically distinct signals must be recognized as equivalent in order to achieve stable perception.

For many years, the prevailing theoretical account of this process was that much of the surface variability in the speech signal was removed via a normalization mechanism (e.g., Ladefoged & Broadbent, 1957; Mullenix et al., 1989; Studdert-Kennedy, 1976). On this view, surface detail manifests as problematic noise for the perceptual system; hence, the role of the normalization mechanism is to create a more pristine signal that can be mapped onto abstract linguistic representations. Under such an account, information regarding the specific acoustic-phonetic information of an utterance is absent from long-term memory. This view has been challenged, however, by findings indicating that listeners do retain in memory many surface characteristics of the speech signal (Church & Schacter; 1994; Nygaard et al., 2000; Palmeri et al., 1993; Schacter & Church, 1992), and that this information can persist in memory for many days (Goldinger, 1996).

One type of surface characteristic that is retained in memory is the phonetic signature associated with individual talkers' voices (e.g., Goldinger, 1998). Goldinger presented listeners with a series of words; for each word, listeners indicated whether the word was old (heard before in the series) or new (not heard before in the series). Results showed that listeners were more accurate at identifying old words when the talker was held constant between subsequent presentations of a given word compared to when the talker varied on each encounter with a given word. This finding indicates that retaining the surface detail of talkers' productions can facilitate recognition memory for individual words.

Findings such as these raise the possibility that talker-specific phonetic variability, previously considered perceptual noise, may be used to customize speech processing for individual talkers, and there is indeed evidence that this is the case. From the domain of spoken word recognition, talker familiarity has been shown to increase intelligibility (Bradlow & Bent, 2008; Nygaard et al., 1994) and decrease processing time (Clarke & Garrett, 2004). These effects hold when listeners learn to identify talkers on the basis of isolated words (Nygaard et al., 1994) or sentences (Nygaard & Pisoni, 1998), and can be achieved even with short periods of exposure (Bradlow & Pisoni, 1999; Clarke & Garrett, 2004). Additional evidence that listeners use talker-specific acoustic-phonetic information to inform perception comes from the domain of talker recognition (e.g., Remez et al., 1981). These studies have shown that even when traditional cues to talker identity (e.g., fundamental frequency, harmonic spectra) have been removed from the signal, listeners can recognize familiar talkers (Remez et al., 1997) as well as learn to identify the voices of unfamiliar talkers (Fellowes et al., 1997), suggesting that talker differences in phonetically-relevant acoustic properties can be sufficient to cue talker identify.

These findings provide evidence that listeners use talker-specific phonetic detail to facilitate higher levels of speech processing, such as 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 /ɛf/ - /ɛs/ continuum. The results showed that listeners adjusted phonetic boundaries so as to include ambiguous tokens within the phonetic category that supported lexical recognition. Relevant to the current work, the lexically-informed boundary adjustment is sometimes applied on a talker-specific basis (Eisner & McQueen, 2005; Kraljic & Samuel, 2005; but see Kraljic & Samuel, 2007) and results from minimal exposure to a talker's productions (Kraljic & Samuel, 2006).[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 & 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

[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.

particular segment in question is unambiguous (i.e., well within a phonetic category) rather than near a boundary. A precursor of such perceptual accommodation is that listeners are sensitive to these kinds of individual talker differences in phonetic properties of speech, and there is recent evidence that this is indeed the case (Allen & Miller, 2004). The property tested by Allen and Miller was voice-onset-time (VOT), which is an articulatory property of stop consonants defined as the time between the release of the stop and the onset of subsequent vocal fold vibration (Lisker & Abramson, 1964). In English, VOT is an important marker of the voicing contrast in that voiced stops are produced with short VOTs and voiceless stops are produced with longer VOTs. Though this relative difference in VOT is sufficient to cue the voicing contrast (Lisker & Abramson, 1970), the absolute VOT produced for a given stop consonant is robustly influenced by contextual factors including speaking rate and place of articulation. In terms of speaking rate, VOTs systematically increase as rate slows (e.g., Kessinger & Blumstein, 1997; Miller et al., 1986; Nagao & de Jong, 2007). In terms of place of articulation, VOTs systematically increase as place moves from front to back in the vocal tract (e.g., Cho & Ladefoged, 1999; Lisker & Abramson, 1964; Volaitis & Miller, 1992). Of particular relevance to the current work, it is also the case that individual talkers differ in their characteristic VOTs, such that some talkers produce longer VOTs than other talkers (Allen et al., 2003).

Allen and Miller (2004) examined whether, when controlling for contextual influences on VOT such as speaking rate and place of articulation, listeners can track talker differences in VOT. In their experiments, listeners participated in training and test phases. In the training phase, listeners learned to identify the voices of two talkers, "Annie" and "Laura". On a single trial during the training phase, 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 phase, 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 the training phase, then they chose the short-VOT variant of Annie's speech at test. Likewise, if they heard Annie produce long VOTs during the training phase, then 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 additional experiments in which listeners were exposed to *town* during the training phases, and then tested on *time.*

That listeners transfered 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. What is not clear from the findings of Allen and Miller is whether the scope of generalization is limited to the voiceless stop presented during training. That is, listeners in their experiments clearly learned how Annie and Laura produced /t/, but they may have also learned how Annie and Laura produced the other two voiceless stops in English, /p/ and /k/. If listeners can transfer information learned in the context of one voiceless stop to other voiceless stops, then they would be informed as to that talker's characteristic productions for an even greater set of lexical items. Such processing would afford faster adaptation to talker-specific phonetic detail as opposed to an adaptation process that requires exposure to each speech segment.

There is strong evidence within the speech perception domain that listeners can transfer information learned in the context of one phonetic segment to similar segments. Some examples of such transfer, focusing on stop consonants, can be found in the literature on selective adaptation and non-native speech sound learning. In the case of selective adaptation, Eimas and Corbit (1973) showed that /pæ/ was an adaptor not only for a labial /bæ/ - /pæ/ continuum, but also for an alveolar /dæ/ - /tæ/ continuum, suggesting that the effects of the adapting stimulus generalized across place of articulation (see also Landahl & Blumstein, 1982; Miller & Eimas, 1976). In the case of non-native speech sound learning, Tremblay et al. (1997) measured cortically evoked responses to auditory stimuli and showed that learning a novel stop voicing contrast at a labial place of articulation transfered to an alveolar place of articulation (see also McClaskey et al., 1983). Given this evidence within the general domain of speech perception, the possibility is raised that transfer of learning across similar phonetic segments will be observed for listeners' accommodation to talker-specific phonetic detail.[2]

In the experiments reported below, we used a slightly modified version of the Allen and Miller (2004) paradigm to examine the scope of generalization involved in listeners' sensitivity to talker differences in VOT. Two experiments were conducted that examined transfer between labial and velar voiceless stops. Experiment 1 examined transfer in a minimal pair context, and Experiment 2 examined transfer in a non-minimal pair context. Within each experiment, we manipulated the direction of transfer: Experiments 1A and 2A examined transfer from /p/ to /k/, and Experiments 1B and 2B examined transfer from /k/ to /p/. This approach not only provided a replication within each experiment, but it also allowed us to confirm that there is not an asymmetry in the direction of transfer.

[2] As stated in Footnote 1, Kraljic & Samuel (2006, 2007) failed to observe talker-specificity in terms of listeners' accommodation of a novel stop voicing contrast. However, results from Kraljic & Samuel (2006) did show generalization across place of articulation. Following exposure to an ambiguous alveolar stop (midway between /d/ and /t/), listeners showed the appropriate lexically-informed boundary adjustment for both an alveolar continuum and a labial continuum. Thus, their findings showed transfer of learning across similar segments, but not on a talker-specific basis. As described in the main text, the current experiments examine whether such transfer will also be observed in cases of talker-specific processing.

In each experiment, listeners were exposed to talkers' characteristic VOTs for one voiceless stop during training phases. Testing took place across two experimental sessions. Session 1 tested performance for the word presented during training. Session 2 was a transfer session, which tested performance for a novel word that began with a different voiceless stop than was presented during training. Table 1.1 provides a summary of the training and test words for the set of experiments.

Table 1.1: Training and test words for Experiments 1 and 2.

		Test	
Experiment	Training	Session 1	Session 2
1 (Minimal pairs)			
1A	*pain*	*pain*	*cane*
1B	*cane*	*cane*	*pain*
2 (Non-minimal pairs)			
2A	*pain*	*pain*	*coal*
2B	*coal*	*coal*	*pain*

In Experiments 1A and 1B, we provided the simplest test of transfer; namely, we examined transfer between words that form minimal pairs. Within each experiment, two groups of listeners participated in training and test phases. During training, listeners heard two female talkers, Annie and Laura, produce one word-initial voiceless stop in the context of a single consonant-vowel-consonant (CVC) word, either *pain* (Experiment 1A) or *cane* (Experiment 1B). 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 (the A-SHORT/ L-LONG training group), and the other group of listeners heard Annie produce relatively long VOTs and Laura produce relatively short VOTs (the A-LONG/L-SHORT training group). The goal of training was for listeners to learn to discriminate the talkers' voices, and, critically, to be exposed to the unique way that Annie and Laura produced the voiceless stop.

In both Experiments 1A and 1B, testing was completed in two sessions. In Session 1, listeners were tested on the word presented during training (*pain* in 1A, *cane* in 1B) and in Session 2, listeners were tested on a novel word (*cane* in 1A and *pain* in 1B). On each trial at test, listeners were presented with a short-VOT and long-VOT variant of either *pain* or *cane* 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), 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. For example, we predicted that listeners in Experiment 1A who heard Annie produce *pain* with short VOTs during training would select the short-VOT variant of *pain* for Annie's speech at test in Session 1. Likewise, we predicted that listeners in Experiment 1B who heard Annie produce *cane* with short VOTs during training would select the short-VOT variant of *cane* for Annie's speech at test in Session 1. The critical question, tested in Session 2, was whether exposure during training would influence which VOT variant was selected for the novel word, and, if so, would it be to the same degree. That is, to what extent does exposure to a talker's characteristic VOTs in the context of *pain* inform the listener as to how that talker produces *cane* (Experiment 1A), and to what extent does exposure to a talker's characteristic VOTs in the context of *cane* inform the listener as to how that talker produces *pain* (Experiment 1B)?

In Experiment 2, we increased the phonological distance between training and Session 2 test words such that they no longer formed minimal pairs, thus increasing the potential difficulty of transfer of talkers' characteristic VOTs across place of articulation. Using the methods outlined for Experiment 1, listeners were exposed to a talker's characteristic VOTs in the context of *pain* (Experiment 2A) or *coal* (Experiment 2B) during training phases. Testing consisted of two sessions; in Session 1, listeners were presented with two VOT variants of the word presented during training (*pain* in 2A, *coal* in 2B) and in Session 2, listeners were presented with two VOT variants of a novel word (*coal* in 2A, *pain* in 2B). In both sessions, listeners were asked to indicate which VOT variant was most representative of that particular talker. The main question, tested in Session 2, was whether exposure during training would influence which VOT variant was selected for the novel word, which differed from the training word in both initial consonant and medial vowel/final consonant, and, if so, would it be to the same degree as that observed for the training word. That is, to what extent does exposure to a talker's characteristic VOTs in the context of *pain* inform the listener as to how that talker produces *coal* (Experiment 2A), and to what extent does exposure to a talker's characteristic VOTs in the context of *coal* inform the listener as to how that talker produces *pain* (Experiment 2B)?

1.2 Experiment 1

Experiment 1 examined transfer of talkers' characteristic VOTs across place of articulation for words that form minimal pairs, focusing on labial and velar voiceless stops. Specifically, we examined transfer from *pain* to *cane* in Experiment 1A and transfer from *cane* to *pain* in Experiment 1B.

1.2.1 Methods

Subjects

Forty subjects were recruited for participation in the experiment. Of the 40 subjects, 20 participated in Experiment 1A and 20 participated in Experiment 1B. For both Experiments 1A and 1B, 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. Four subjects were replaced for this reason.

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 involved eight steps.

Step 1: Acquiring matched voiced-initial tokens from two talkers. 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.

Two talkers, referred to as Annie and Laura, were selected. The selected talkers had perceptually distinct voices and roughly comparable overall word durations. One repetition of *bane* and *gain* was selected from each talker such that VOT for a given word was approximately matched across the talkers. For Annie, VOTs of the selected *bane* and *gain* tokens were 0 ms and 17 ms, respectively. For Laura, VOTs of the selected *bane* and *gain* tokens were 0 ms and 19 ms, respectively. The four selected tokens were first 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.

Step 2: Creating synthesized versions of the matched voiced-initial tokens. 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 is a numeric table that displays peak amplitude, fundamental frequency (F_0), and formant frequencies/bandwidths for each frame of the analyzed token. These data, along with the residual excitation, were used to create a synthesized version of each selected token. In order to ensure that the synthesized token preserved the release burst of the original token, the first impulse marker, as calculated by the LPC analysis, was deleted as necessary (thus increasing the length of the first frame). Also when necessary, a scaling factor of 0.25 was applied to the final frames of the synthesized token in order to yield a more natural word offset.

Step 3: Creating VOT series based on the synthesized voiced-initial tokens. 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*. The first token of each series used the parameters of the voiced-initial tokens as described in Step 2. Additional steps were created by successively converting voiced frames to voiceless frames by setting the excitation parameter to a noise source, setting the F_0 parameter to zero, and applying a scaling factor of 0.15 to the peak amplitude parameter. (An exception to this procedure concerns the second step of the series; in order to preserve the release burst, the scaling factor applied to the peak amplitude parameter was adjusted when necessary.) Each series consisted of 40 tokens, ranging in VOT from 0 ms to 200 ms for the labial-initial series and approximately 20 ms to 220 ms for the velar-initial series. For all series, the step size in VOT was 4 to 5 ms, which corresponds to the duration of each successive pitch period. This procedure yielded a pool of tokens that were matched on overall duration and differ in word-initial VOT; of these tokens, some specify the voiced-initial endpoint of a particular series, and many others (spanning a wide range of VOTs) specify the voiceless-initial tokens of a particular series.

Step 4: Selecting stimuli from the VOT series to be presented during training. Five tokens were selected from each VOT series to serve as training stimuli, including one voiced-initial token and four voiceless-initial tokens. Of the four voiceless tokens, two were selected such that they had relatively short VOTs and two were selected such that they had relatively long VOTs. The VOT values of the selected tokens are shown in Table 1.2 for each training group. The particular tokens were selected as follows: (1) the first step of each series was the voiced token; (2) two tokens from the short-VOT voiceless region of each series were selected such that they were two steps apart on the continuum (to simulate naturally occurring within-talker variability) and VOT for a particular word was closely matched across talkers; (3) two tokens from the long-VOT voiceless region of each series were likewise selected such that they were two steps apart on the continuum and VOT for a particular word was closely matched across talkers; (4) the short-VOT and long-VOT voiceless tokens were selected so as 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. As stated previously, place of articulation yields a systematic influence on VOTs in speech production, such that VOTs for labial stops are shorter than VOTs for velar stops, at a given rate of speech. As shown in Table 1.2, the stimuli selected for use in the current experiments are in accord with how these values pattern in the production of natural speech.

Step 5: Selecting stimuli from the VOT series to be presented at test. Both training groups were presented with the same test stimuli. Two tokens were selected from each VOT series for use during test, 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 1.3. The particular tokens were

Table 1.2: VOT values (ms) of the *bane*/*pain* and *gain*/*cane* training stimuli.

Training Group: A-SHORT / L-LONG

		pain			*cane*	
Talker	*bane*	Token 1	Token 2	*gain*	Token 1	Token 2
Annie	0	60	69	17	83	92
Laura	0	155	164	19	178	187

Training Group: A-LONG / L-SHORT

		pain			*cane*	
Talker	*bane*	Token 1	Token 2	*gain*	Token 1	Token 2
Annie	0	155	165	17	177	186
Laura	0	60	68	19	83	92

Table 1.3: VOT values (ms) of the *pain* and *cane* test stimuli.

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

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.

Step 6: Eliminating potential amplitude-based confound. Two amplitude variants were created for the selected training and test tokens in order to eliminate a potential confound that results from the synthesis techniques used to create the VOT series. With these techniques, tokens with shorter VOTs have higher overall amplitude (measured in terms of RMS level) than tokens with longer VOTs. Thus, overall amplitude of the short-VOT training and test tokens is higher than the overall amplitude of the long-VOT training and test tokens. This raises the possibility that performance at test could reflect sensitivity to the amplitude difference between the short-VOT and long-VOT tokens, rather than sensitivity to VOT per se.

This potential amplitude-based confound was eliminated as follows. A high- and low-amplitude variant was made for all tokens drawn from the *bane/ pain* and *gain/cane* sets, one corresponding to the mean RMS amplitude of the short-VOT tokens and the other corresponding to the mean RMS amplitude of the long-VOT tokens. At presentation, amplitude of the high and low variants was 67 dB SPL and 65 dB SPL, respectively. As described below, the multiple amplitude variants were presented during training and test, thus ensuring that subjects' performance could not be attributed to the amplitude difference of the short-VOT and long-VOT tokens that resulted from the synthesis techniques used to generate the stimuli.

Step 7: Constructing training stimulus lists. For each of the *bane/pain* and *gain/cane* stimulus sets, separate training lists were created for the A-SHORT/L-LONG training group and the A-LONG/L-SHORT training group. Each list contained both amplitude variants of each training stimulus. The training lists for the A-SHORT/L-LONG training group contained Annie and Laura's voiced-initial tokens (*bane* in Experiment 1A, *gain* in Experiment 1B), Annie's short-VOT voiceless-initial tokens (*pain* in Experiment 1A, *cane* in Experiment 1B), and Laura's long-VOT voiceless-initial tokens (*pain* in Experiment 1A, *cane* in Experiment 1B). The training lists for the A-LONG/L-SHORT training group contained, in analogous fashion, Annie and Laura's voiced-initial tokens, Annie's long-VOT voiceless-initial tokens, and Laura's short-VOT voiceless-initial tokens. In each training list, an extra voiced-initial 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 voiced-initial tokens X 2 voiceless-initial tokens X 2 amplitude levels) in randomized order. Sixteen such lists were created for each stimulus set to be presented to the listeners during training.

In addition to the training lists, stimulus lists were created for use during a familiarization phase and a practice phase (details on the phases of the experiment are provided in the following procedure section). The familiarization list was created following the method outlined for creation of the training lists. The practice phase was blocked by talker's voice, thus four lists were made for each stimulus set, one for each talker for each training group. Each practice list consisted of three randomized blocks of the eight stimuli to be used during training (2 voiced-initial tokens X 2 voiceless-initial tokens X 2 amplitude variants), yielding 24 practice trials for each talker.

The *bane/pain* familiarization, practice, and training lists were used in Experiment 1A; and the *gain/cane* familiarization, practice, and training lists were used in Experiment 1B.

Step 8: Constructing test stimulus lists. 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 1500 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, four for each talker for each stimulus set. One additional test list was created for each talker, for each stimulus set, for use during the practice phase of the experiment following the same procedure. In Experiment 1A, the *pain* test lists were used in Session 1 and the *cane* test lists were used in Session 2. In Experiment 1B, the *cane* test lists were used in Session 1 and the *pain* test lists were used in Session 2 (see Table 1.1).

Procedure: Session 1

Experiment 1A. Twenty subjects participated in Experiment 1A. Half were assigned to the A-SHORT/L-LONG training group and half were assigned to the A-LONG/L-SHORT training group. All testing took place in a sound-attenuated booth, and auditory stimuli were 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 condition. At test, subjects in both training conditions were presented with the *pain* test lists.

Prior to training, subjects completed a brief familiarization phase. The purpose of the familiarization phase was for listeners to learn to identify each talker's voice. During familiarization, one set of training stimuli 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.

After familiarization, subjects completed a brief practice phase in order to be exposed to the training and test tasks. The practice phase was blocked by voice, with the order of the voices counter-balanced within each training group. For each voice, subjects completed a practice training phase and a practice test phase (using the stimulus lists generated for the practice phases). During training practice, subjects were presented with three random orderings of the eight training stimuli. On each trial, they were asked to identify the initial consonant. They indicated their response by pressing a button labeled B or P on a response keypad. No feedback was provided. During test practice, subjects were presented with the practice test set for that talker. 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.

Following familiarization and practice, the experiment proper began with the alternation between training phases and test phases. In each training phase, subjects were presented with one training set, with the order of the training sets 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 sets for one of the talkers. The order of presentation for the test sets was determined randomly for each subject, with the constraint that no more than three tests sets of the same talker were presented in a row. Instructions during the test phase were identical to those described above for the practice test phase. On each test trial, subjects indicated which of the two VOT variants presented on each trial was most representative of that talker's voice. The delay between the two VOT variants on a given trial was 750 ms. The pause between trials was 2000 ms, timed from the button response.

Overall, the sequence of the phases was as follows: familiarization phase, 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.[3]

Experiment 1B. All procedural details outlined for Experiment 1A were followed in 1B, save that *gain/cane* training stimuli and the *cane* test stimuli were presented in the appropriate phases.

[3] 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.

Procedure: Session 2

Following a brief break after Session 1, subjects completed Session 2, which was a transfer session. In Session 2, subjects completed an additional eight alternations of training phases and test phases. Stimuli during training remained the same as presented in Session 1, but the test stimuli changed. In Experiment 1A, where listeners were trained using the *bane/pain* stimuli, listeners were tested on the *cane* test sets. In Experiment 1B, where listeners were trained using the *gain/cane* stimuli, listeners were tested on the *pain* test sets. The procedural details for the training and test phases followed those outlined for Session 1.

1.2.2 Results

Experiment 1A

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 was adopted to indicate high performance during training. Two subjects were replaced because they did not reach the criterion for talker identification.

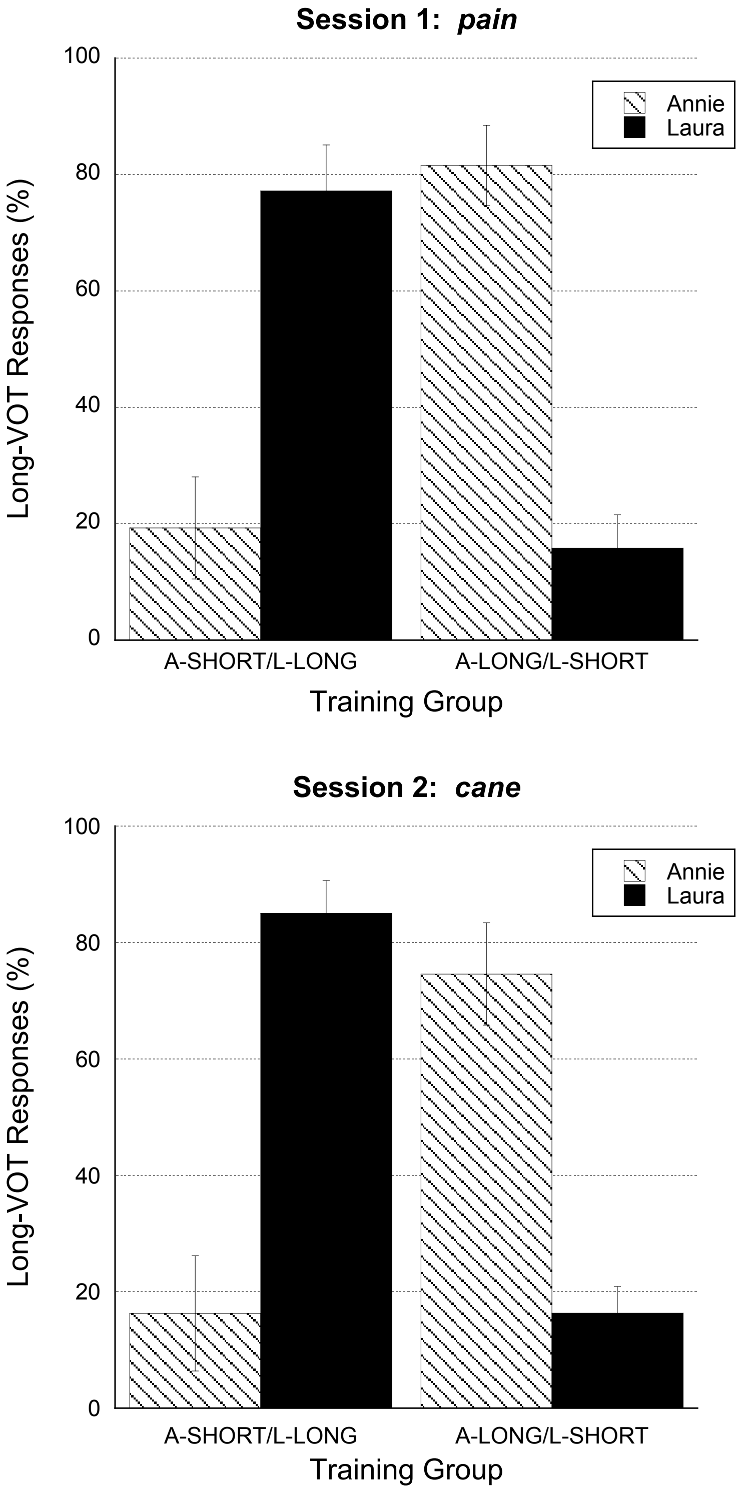
For the twenty subjects included in the experiment, performance during training across both sessions was near ceiling for both talker identification (95%) and phonetic identification (99%).

Test. 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 short-VOT and long-VOT responses must sum to 100, quantifying performance in both terms is redundant.] For each subject, mean percent long-VOT responses was calculated for a given talker separately for each session. Figure 1.1 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.

Consider first performance during Session 1. Two sets of analyses were performed. In the primary analysis, the percentage of long-VOT responses was submitted to ANOVA with the between-subjects factor of training group and the within-subjects factor of talker. Results of the ANOVA showed no main effect of either training group [$F(1, 18) < 1$] or talker [$F(1, 18) < 1$], but a significant interaction between training group and talker [$F(1, 18) = 42.06$, $p < .001$].

The purpose of the secondary set of analyses was two-fold. 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

Figure 1.1: Mean percent long-VOT responses for the test phases in Experiment 1A 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.



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 Figure 1.1. 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. To examine the statistical significance of this pattern, the same analyses described for Session 1 were performed for Session 2. The results from the primary analysis showed no main effect of talker [F(1,18) $<$ 1] or training group [F(1,18) = 1.02, p = .327], but a significant interaction between these two factors [F(1,18) = 46.27, p $<$ .001]. 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, and no main effect of talker or training group [F(1,18) $<$ 1 in both cases]. Critically, there was no effect of session [F(1, 18) $<$ 1] and no interaction between session and talker [F(1, 18) = 1.55, p = .229] or session and training group [F(1, 18) = 2.26, p = .150]. Moreover, the 3-way interaction between talker, training group, and session was not significant [F(1,18) $<$ 1]. These results indicate that performance at test was as robust for the novel word as it was for the training word.

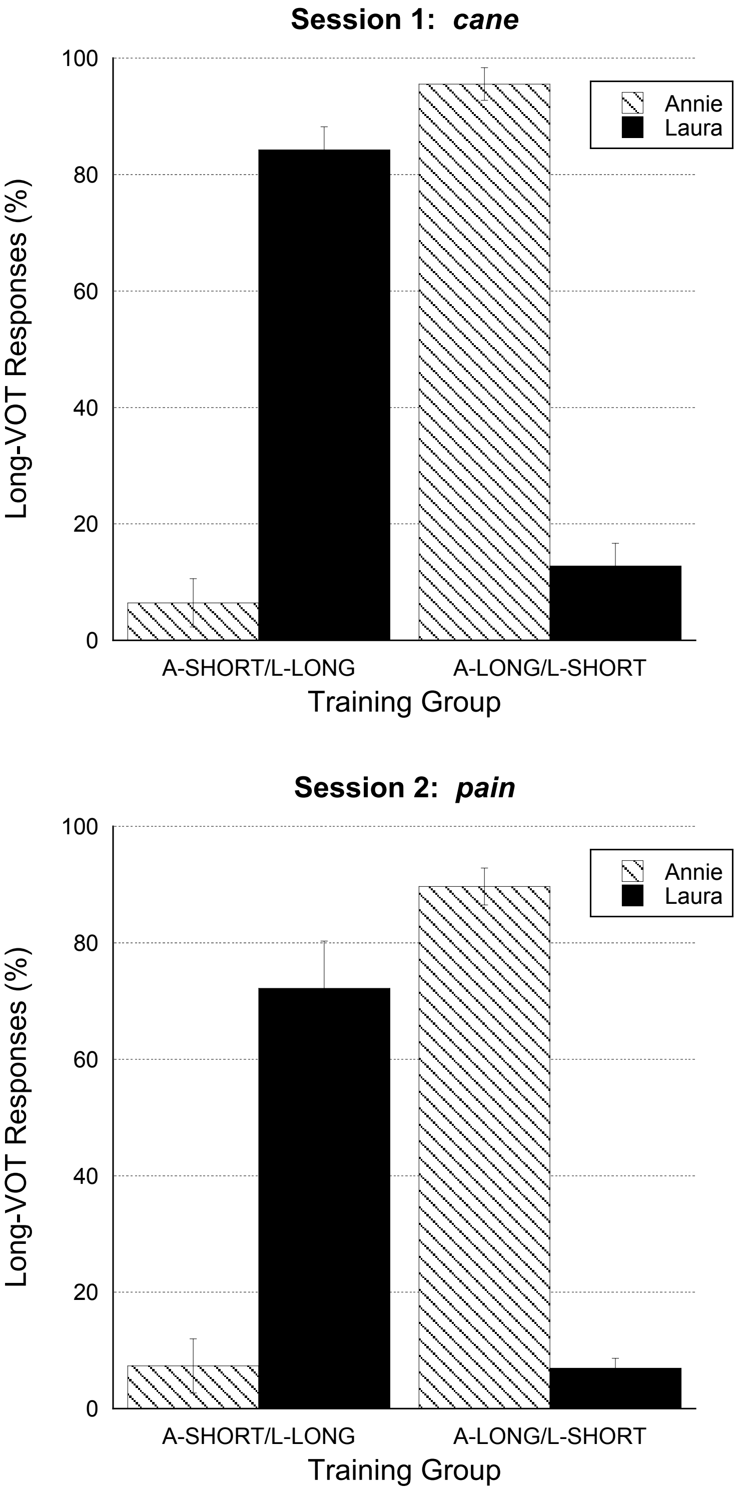
Experiment 1B

Training. Performance during training was analyzed as outlined for Experiment 1A. Two subjects were replaced because they did not reach criterion on talker identification. For the twenty subjects included in the experiment, mean percent talker identification and phonetic identification across sessions were near ceiling (98% and 99%, respectively).

Test. Figure 1.2 shows percent long-VOT responses for each training group separately for each talker; Session 1 data are shown in the top panel and Session 2 data are shown in the bottom panel. The key pattern of results found for Experiment 1A was also observed here, and this was confirmed statistically following the analyses described previously. Results from the primary analysis for Session 1 revealed a strong interaction between talker and training group [$F(1,18) = 372.04$, $p < .001$] and no main effect of talker [$F(1,18) < 1$]. Unlike in Experiment 1A, the main effect of training group was significant [$F(1,18) = 7.32$, $p = .015$], and resulted from fewer long-VOT responses in the A-SHORT/L-LONG training group (45%) compared to the A-LONG/L-SHORT training group (54%).

Results from the primary analysis for Session 2 showed no main effect of talker [$F(1,18) = 2.59$, $p = .125$], a marginal effect of training group [$F(1,18) = 3.76$, $p = .068$], and an interaction between talker and training group [$F(1,18) = 177.03$, $p < .001$]. In both Session 1 and Session 2, results from the

Figure 1.2: Mean percent long-VOT responses for the test phases in Experiment 1B 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.



secondary analyses confirmed that the percentage of long-VOT responses for each talker for each training group was different from chance and that the interaction was due to the predicted pattern of results.[4]

These analyses indicate that listeners used experience during training to guide performance at test for the training word *cane*, and transfered this information to the novel word *pain*. As for Experiment 1A, an additional ANOVA was conducted in order to examine the strength of transfer. The percentage of long-VOT responses was submitted to an ANOVA with the factors talker, training group, and session. Results showed no main effect of talker [F(1,18) = 1.88, p = .187], and, as observed in Session 1, a main effect of training group [F(1,18) = 10.50, p = .005]. In addition, the interaction between talker and training group was confirmed [F(1,18) = 342.34, p < .001]. There was a marginal effect of session [F(1,18) = 4.18, p = .056], but, critically, no interaction between session and talker [F(1,18) = 1.54, p = .230] or session and training group [F(1,18) < 1]. Moreover, the 3-way interaction between talker, training group, and session was not reliable [F(1,18) = 1.59, p = .223].

Viewed together, results from Experiment 1 indicate that listeners can track a talker's characteristic VOTs, and, moreover, can transfer this information to a word that begins with a different voiceless stop. Transfer across place of articulation is not contingent on a particular direction of transfer. As described

[4] The locus of the main effect of training group is not known. That it was significant in Session 1 and marginally significant in Session 2 suggests that it may be a characteristic of this group of listeners. However, it was not observed in Session 2 of Experiment 1A, which used the same stimuli, suggesting that it is not a characteristic of the *gain*/*cane* stimulus set.

previously, place of articulation exhibits a systematic contextual influence on VOTs in speech production, simulated in the current experiments, such that VOTs for labial stops are shorter than VOTs for velar stops. Thus, in Experiment 1A, listeners transfered from labial-initial *pain* to the slightly longer VOTs in velar-initial *cane*, and in Experiment 1B listeners transfered from *cane* to the slightly shorter VOTs in *pain*. Critically, regardless of the direction of transfer, sensitivity to characteristic VOTs was not lessened for the novel word compared to the training word. These data demonstrate not only that transfer is possible across place of articulation, but also that it is robust in that the strength of sensitivity to the fine-grained differences in production across the two talkers was not mediated by which voiceless stop was presented during training.

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. The results indicated robust transfer. In Experiment 2, we examined whether such transfer is limited to this constrained environment or if transfer would also be observed between words that are phonologically less similar.

1.3 Experiment 2

Experiment 2 examined transfer of talkers' characteristic VOTs across place of articulation for words that do not form minimal pairs. Specifically, we examined transfer from *pain* to *coal* in Experiment 2A and transfer from *coal* to *pain* in Experiment 2B.

1.3.1 Methods

Subjects

Forty different subjects were recruited for participation in the experiment following criteria outlined for Experiment 1. Twenty of the subjects participated in Experiment 2A and twenty participated in Experiment 2B. In both Experiments 2A and 2B, 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. Two subjects were replaced because they did not reach criterion performance during training.

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. As shown in Table 1.1, the *bane/pain* and *goal/coal* sets were used in both Experiment 2A and 2B. Preparation of the *goal/coal* set followed the procedures outlined for Experiment 1, and is summarized below.

Stimuli. One token of *goal* was selected for each talker from the recordings described for Experiment 1. VOTs of the selected *goal* tokens were 25 ms for Annie and 29 ms for Laura. Both tokens were trimmed to 568 ms in duration (with appropriate amplitude ramping at offset) 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. Five tokens were selected from each series to be presented during training. The VOTs of the selected training tokens are shown in Table 1.4.

Table 1.4: VOT values (ms) of the *goal/coal* training stimuli.

Training Group: A-SHORT / L-LONG			
		coal	
Talker	*goal*	Token 1	Token 2
Annie	24	83	92
Laura	29	178	190

Training Group: A-LONG / L-SHORT			
		coal	
Talker	*goal*	Token 1	Token 2
Annie	24	180	189
Laura	29	84	92

In addition, two tokens were selected from each series to be presented during test, a short-VOT *coal* token and a long-VOT *coal* token. The VOTs of the selected test tokens are shown in Table 1.5. For each of the selected training and test tokens, a high- and low-amplitude variant were generated in order to eliminate a potential amplitude-based confound. Generating the amplitude variants involved two steps. First, two amplitude variants were made for all tokens selected from the *goal/coal* set, one corresponding to the mean RMS level of the short-VOT tokens and one corresponding to the mean RMS level of the long-VOT tokens. As in Experiment 1, these RMS levels were offset by 2 dB. Second, due to differences in intrinsic vowel loudness, RMS amplitude of the high and low variants was increased by 4 dB in order to match loudness of the *goal*/*coal* tokens to the loudness of the *bane*/*pain* tokens. At presentation, amplitude of the high and low variants for the *goal*/*coal* set was 71 dB SPL and 69 dB SPL, respectively. As reported in Experiment 1, amplitude of the high and low variants for the *bane*/*pain* set was 67 dB SPL and 65 dB SPL, respectively.

Table 1.5: VOT values (ms) of the *coal* test stimuli.

	coal	
Talker	Short-VOT	Long-VOT
Annie	87	185
Laura	88	182

Constructing training stimulus lists. The *bane*/*pain* training lists used in Experiment 1A were used in Experiment 2A. For Experiment 2B, sixteen training lists using the *goal*/*coal* stimulus set were constructed for each training group using both amplitude variants of the selected training stimuli. For the A-SHORT/L-LONG training group, the training list contained Annie and Laura's *goal* tokens, Annie's short-VOT *coal* tokens, and Laura's long-VOT *coal* tokens. For the A-LONG/L-SHORT training group, the training list contained Annie and Laura's *goal* tokens, Annie's long-VOT *coal* tokens, and Laura's short-VOT *coal* tokens. Each training list consisted of 16 tokens in a randomized order (2 talkers

X 2 voiced-initial tokens X 2 voiceless-initial tokens X 2 amplitude levels). As for Experiment 1, the *goal/coal* training stimuli were also used to generate familiarization and practice lists.

Constructing test stimulus lists. Multiple *coal* test lists were created for each talker following the procedures outlined in Experiment 1. As shown in Table 1.1, these lists were used during Session 2 in Experiment 1A and Session 1 in Experiment 2B.

Procedure

The overall procedure for Experiment 2 followed that outlined for Experiment 1. In brief, subjects in both Experiment 2A and Experiment 2B participated in two sessions. Session 1 began with a brief familiarization phase. Following familiarization, subjects completed the practice phase, and then began to alternate between training and test phases to the completion of eight test phases, four for Annie's voice and four for Laura's voice. Session 2 consisted of an additional eight alternations between training and test phases; the training phases used the same stimuli as Session 1, but the test stimuli differed. The *bane/pain* training lists were used in Experiment 2A and the *goal/coal* training lists in Experiment 2B. For Experiment 2A, the *pain* test lists were used in Session 1 and the *coal* test lists were used in Session 2. For Experiment 2B, the *coal* test lists were used in Session 1 and the *pain* test lists were used in Session 2 (see Table 1.1).

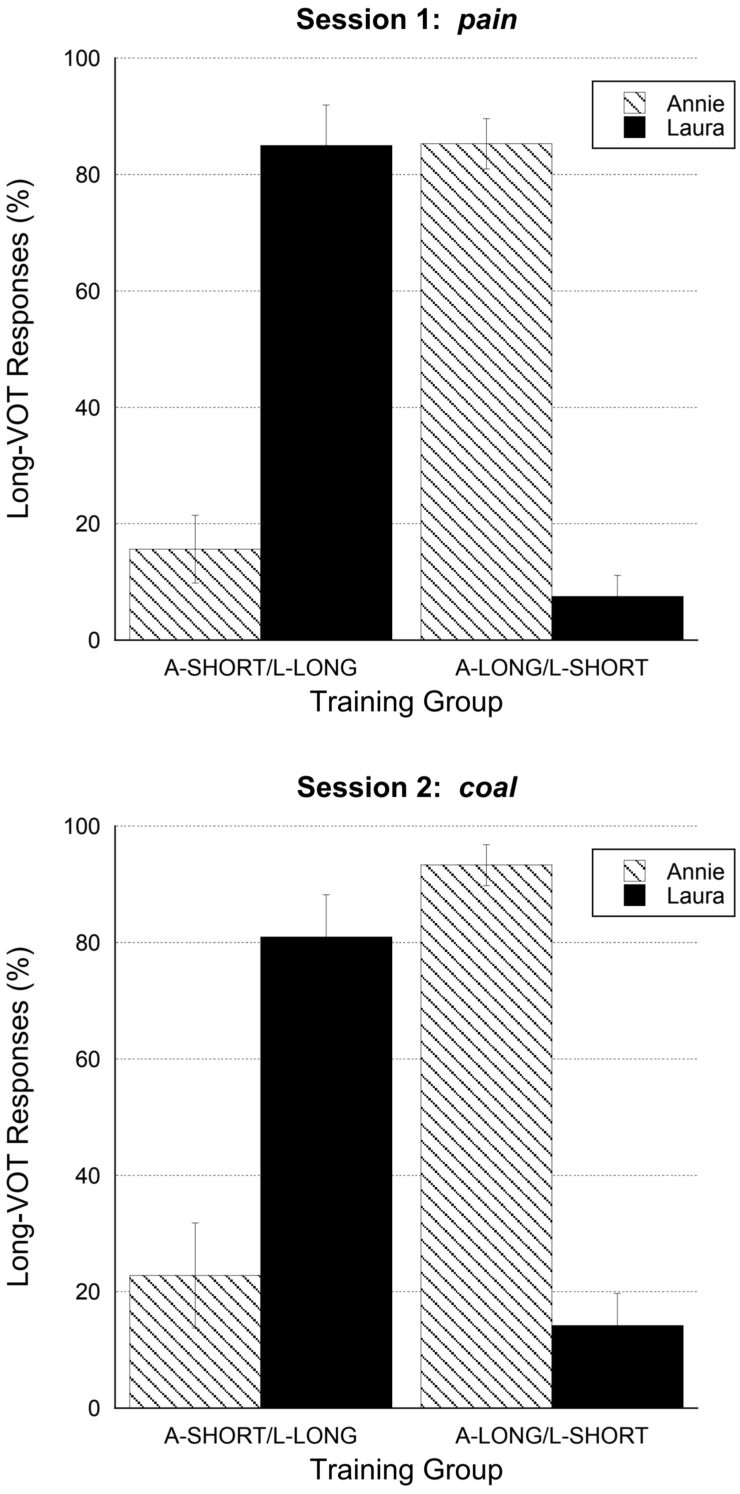
1.3.2 Results

Experiment 2A

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

Test. Figure 1.3 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 Experiments 1A and 1B. 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 revealed no main effect of talker [$F(1,18) < 1$] or training group [$F(1,18) = 1.35$, $p = .260$], but a significant interaction between talker and training group [$F(1,18) = 118.59$, $p < .001$]. The same pattern was observed in Session 2; ANOVA showed no main effect of talker [$F(1,18) = 1.37$, $p = .258$] or training group [$F(1,18) < 1$], but a strong interaction between these two factors [$F(1,18) = 58.49$, $p < .001$]. For both Session 1 and Session 2, results from the secondary analyses confirmed that the percentage of long-VOT

Figure 1.3: Mean percent long-VOT responses for the test phases in Experiment 2A 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.



Session 1: *pain*

Session 2: *coal*

responses for each talker in each training group was different from chance, and that the interactions reported above were due to the predicted pattern of results. That is, in both sessions, 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. In addition, 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.

As in Experiments 1A and 1B, 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 [$F(1,18) = 96.63$, $p < .001$], and no main effect of training group [$F(1,18) < 1$] or talker [$F(1,18) = 1.03$, $p = .323$] was observed. In addition, there was no main effect of session [$F(1,18) = 3.39$, $p = .082$], nor was there an interaction between session and talker [$F(1,18) < 1$] or session and training group [$F(1,18) = 1.43$, $p = .247$]. Furthermore, the 3-way interaction between talker, training group, and session was not significant [$F(1,18) < 1$], indicating that performance at test was as robust for the novel word as it was for the training word.

Experiment 2B

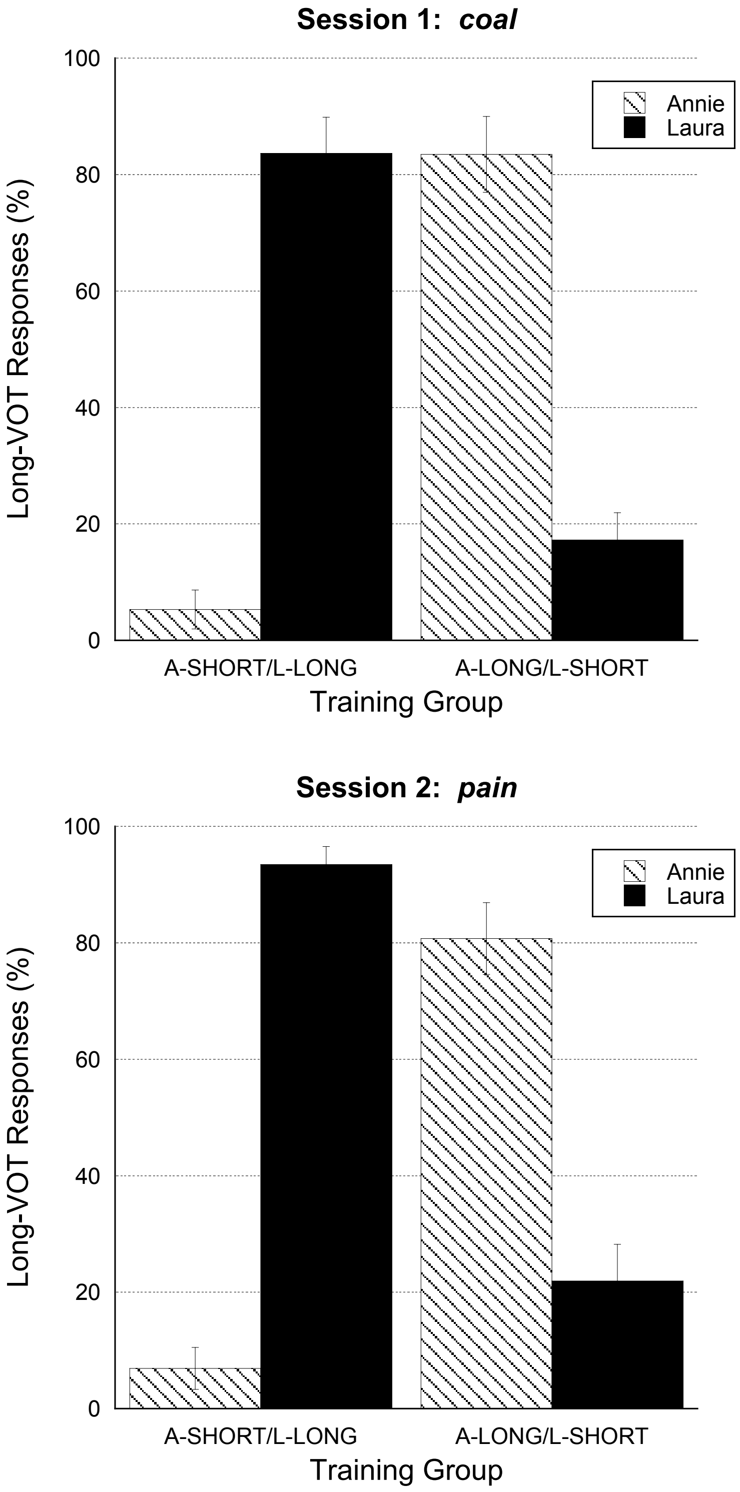
Training. Performance during training was analyzed as outlined previously. No subjects were replaced for sub-criterion performance.

Performance during training was near ceiling for both talker identification (99%) and phonetic identification (99%).

Test. Figure 1.4 shows percent long-VOT responses for each training group separately for each talker; Session 1 data are shown in the top panel and Session 2 data are shown in the bottom panel. The performance observed here is similar to that seen in Experiment 2A. Results of the primary analysis for Session 1 showed no main effect of talker [F(1,18) $<$ 1] or training group [F(1,18) = 2.75, p = .114], and confirmed the interaction between talker and training group [F(1,18) = 117.48, p $<$ .001]. For Session 2, results of the primary analysis showed a main effect of talker [F(1,18) = 4.63, p = .045], with fewer long-VOT responses for Annie's voice (44%) compared to Laura's voice (58%), no main effect of training group [F(1,18) $<$ 1], and, critically, a significant interaction between talker and training group [F(1,18) = 127.08, p $<$. 001].[5] Results from the secondary analyses confirmed that, in each session, the percentage of long-VOT responses for each talker for each training group was different from chance and that the interaction between talker and training group was due to the predicted pattern of results. To assess the strength of transfer, an additional ANOVA with the factors of talker, training group, and session was performed on percent long-VOT responses. Results showed no main effect of talker [F(1,18) = 2.64, p = .121] or training group [F(1,18) = 2.23, p = .153], and

[5] The locus of the main effect of talker reported here is not known. That it was observed in Session 2 but not Session 1 suggests that it is not consistent for this group of listeners. Furthermore, it was not observed in any other experiment that used the *bane/pain* stimulus set; thus, it is not consistent with these particular stimuli.

Figure 1.4: Mean percent long-VOT responses for the test phases in Experiment 2B 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.



Session 1: *coal*

Session 2: *pain*

confirmed the interaction between talker and training group [F(1,18) = 139.82, $p < .001$]. There was no main effect of session [F(1,18) = 2.29, p = .148], and session did not interact with either talker [F(1,18) = 2.80, p = .111] or training group [F(1,18) = 1.14, p = .300]. Moreover, the interaction between talker, training group, and session was not significant [F(1,18) < 1], indicating no difference in performance for the novel word compared to the training word.

Collectively, results from Experiment 2 further indicate that listeners can transfer information learned about a talker's characteristic VOTs from one voiceless stop to a different voiceless stop, and, critically, that such transfer is not constrained to a minimal pair context. As in Experiment 1, transfer across place of articulation was not contingent on a particular direction of transfer. Moreover, sensitivity to characteristic VOTs was as strong for the novel word as it was for the training word.

1.4 Discussion

The acoustic signal of speech is highly variable. Many factors influence the precise acoustic-phonetic information produced for individual speech segments such as phonetic context (Delattre et al., 1955) and speaking rate (Miller, 1981). Yet another source of variability in the speech signal, and that considered in the current experiments, stems from individual talker differences in speech production (e.g., Allen et al., 2003). Despite such variability in the acoustic-phonetic input, listeners reliably extract linguistic units from the speech signal.

A large body of research within the domain of speech perception has examined how listeners accommodate for the lack of invariance between the acoustic signal and linguistic percept. Contrary to traditional normalization accounts of speech perception, there is much evidence indicating that listeners accommodate for 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 retained and 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) and talker identification (Remez et al., 1981), 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 stop consonants, Allen and Miller (2004) showed that listeners could learn that one talker produced characteristically short VOTs and that a different talker produced characteristically long VOTs. The goal of the current work was to examine the scope of generalization underlying such listener 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 one voiceless stop consonant, either /p/ or /k/, in the context of a word (e.g., *pain* or *cane*). Sensitivity to talkers' characteristic VOTs was assessed for the word presented during training as well as for a novel word that began with a different voiceless stop than presented during training. Across the two experiments, we manipulated the phonological distance between the training and novel words; specifically, the words formed minimal pairs (*pain* and *cane*) in Experiment 1 but did not form 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, replicating earlier findings (Allen & Miller, 2004), but also for the novel word. Moreover, for both the minimal pair and non-minimal pair cases, 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. In other words, complete transfer of learning was obtained. That transfer across place of articulation was observed indicates that listeners do not require exposure to each individual segment in order to "tune in" to a talker's phonetic signature; rather, there is generalization across similar segments. One striking aspect of the talker-specificity effects reported for higher levels of processing, described above, is that the processing advantage achieved by talker familiarity also generalizes to novel items (Nygaard & Pisoni, 1998). Such broad scope of generalization, at both the segmental and lexical levels, potentially affords more efficient adaptation to talker-specific phonetic detail compared to a learning process that operates in a segment-by-segment fashion, and may in fact underly other findings indicating that adaptation to this type of variability in the speech signal is a rapid process (Clarke & Garrett, 2004).

As stated previously, the current findings indicate that listeners do not require exposure to each voiceless stop in order to learn characteristic VOTs; rather, there is transfer across place of articulation. However, an additional issue that will need to be examined in order to fully describe how listeners accommodate this type of talker-specific phonetic detail concerns variation 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. As described below, VOT is robustly influenced by variation in speaking rate (e.g., Miller et al., 1986); moreover, talkers frequently alter their speaking rates (Miller et al., 1984). Thus, a complete examination of sensitivity to this acoustic-phonetic property of speech must consider transfer of learning when speaking rate varies.

The effect of speaking rate on VOT has been examined extensively, both in the production and perception domains. In terms of speech production, it has long been known that VOT is influenced by speaking rate, with VOTs systematically increasing as speaking rate slows (Miller et al., 1986). Findings from the perception domain indicate that listeners take speaking rate into account when processing VOT (Miller, 1981). Not only does the voicing boundary shift to longer VOTs as speaking rate slows (Summerfield, 1981), but so too do those members of the voiceless stop category rated most prototypical (Miller & Volaitis, 1989). Given that listeners are highly sensitive to the effect of speaking rate on VOT, and that talkers not only differ in their characteristic VOTs (Allen et al., 2003) but also frequently alter their rates of speech (Miller et al., 1984), future work will need to determine how listeners accommodate for talker differences in VOT with respect to the influence of speaking rate on VOT. One key question concerns the type of exposure listeners will require in order to track talker differences in VOT across variation in speaking rate. In the current experiments, exposure to one CVC word beginning with a voiceless stop was enough to inform listeners as to how the talkers produced a different voiceless stop. Will such minimal exposure afford transfer to words produced at a novel speaking rate? And, if so, will transfer be observed across a simultaneous change in speaking rate and place of articulation?

Future work is also needed to explicate the representational consequences of sensitivity to talker differences in phonetic properties of speech. One central issue concerns the mechanism of transfer. Though the current findings demonstrate robust transfer in learning talker-specific phonetic detail, they do not identify the mechanism by which transfer is obtained. 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. 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 VOT that was presented during training. Determining the mechanism underlying the observed transfer is necessary in order to fully describe the scope of generalization involved in learning talker-specific phonetic detail. As a case in point, if transfer operates via phonetic features, then the scope of generalization may be more limited than if it were to operate along dimensions of acoustic similarity.

A second central issue concerns phonetic category representation. As reviewed previously, there is evidence indicating that listeners accommodate ambiguous idiosyncratic productions (e.g., a sound midway between /s/ and /f/) by adjusting phonetic category boundaries (Norris et al., 2003). Unlike the ambiguous productions examined previously, the current work examines sensitivity to talker-specific productions that are well-defined, or unambiguous, category members. The current findings demonstrate that listeners can track talker differences in phonetic properties of speech even when the productions are clear exemplars, but it is not known whether this tracking fundamentally alters the mapping from speech signal to segmental representation.

One way that listeners may adjust the mapping process in order to take into account talker-specific phonetic detail for clearly defined category members is to alter the internal structure of phonetic categories in line with a talker's characteristic productions. It has long been known that phonetic categories are marked not only by boundaries between them, but that a given category exhibits an internally graded structure in that not all members are considered equally good members (e.g., Kuhl, 1991; Samuel, 1982). Such graded structure has been demonstrated for many different speech sounds, including both consonants (Volaitis & Miller, 1992) and vowels (Kuhl, 1991). Furthermore, the internal structure of phonetic categories has been shown to be highly sensitive to contextual influences in speech production (e.g., Allen & Miller, 2001). As an example, consider the case of word-initial VOT specifying a voiceless stop consonant. In line with the contextual influence of speaking rate on VOT in speech production, VOTs corresponding to the highest rated exemplars of a particular voiceless stop are shorter for fast speaking rates compared to slow speaking rates (Miller & Volaitis, 1989). Given that the internal structure of phonetic categories is highly tuned to contextual influences in speech production, the possibility is raised that listeners might customize the internal category structure for individual talkers. That is, talker identity may act as a contextual influence on phonetic category representation such that the most prototypical members of a particular phonetic category shift along acoustic-phonetic space in order to be centered on characteristic productions of individual talkers.

The present findings indicate that accommodating talker differences in speech production does not entail exposure to each individual speech segment. In the case of word-initial VOT, learning how a talker implements one voiceless stop informs the listener as to how that talker produces a different voiceless stop, demonstrating broad scope of generalization in adjusting to talker differences in phonetic properties of speech. Future work is aimed at further explicating the nature of listener adaptation to talker-specific phonetic detail.

Chapter 2

Talker-specific phonetic detail in speech production

2.1 Introduction

The past fifty years of research in speech acoustics have yielded substantial information on the acoustic parameters that specify individual speech segments. One consistent finding in this domain is that there is considerable variability in the acoustic-phonetic information produced for individual consonants and vowels, such that there is no one-to-one mapping between the acoustic signal and speech segment. Many sources of acoustic-phonetic variability have been examined, including variability that results from differences in pronunciation across individual talkers. Talker differences have been observed for a host of speech sound classes including vowels (Hillenbrand et al., 1995; Peterson & Barney, 1952), fricatives (Newman et al., 2001), stops (Allen et al., 2003; Byrd, 1992; Zue & Laferriere, 1979), and liquids (Espy-Wilson et al., 2000; Hashi et al., 2003). The goal of the current work is to characterize further such talker differences.

Talker differences in phonetically-relevant properties of speech are theoretically important in terms of describing how listeners recover the segmental structure of language during comprehension. Early accounts of speech perception posited that perceptual constancy for spoken language was achieved via a normalization mechanism, such that variability in the speech signal was discarded early in the perceptual process in order to map the speech signal onto abstract prelexical representations (e.g., Studdert-Kennedy, 1976). Under this account, information about the specific phonetic details of an utterance is absent from long-term memory. However, more recent findings indicate that listeners retain fine-grained information about how a talker implements speech segments (Goldinger, 1998; Palmeri et al., 1993) and that this information can persist in long-term memory for many days (Goldinger, 1996). These data challenge strict normalization accounts of speech perception and raise the possibility that instead of discarding talker-specific acoustic-phonetic information, listeners retain this information and use it to facilitate perception.

In support of this alternative account, there is now evidence that familiarity with a particular talker's speech can facilitate subsequent processing. Talker familiarity has been shown to increase comprehension (Bradlow & Bent, 2008; Nygaard et al., 1994) and decrease processing time (Clarke & Garrett, 2004). For example, Nygaard et al. (1994) trained listeners to identify the voices of 10 talkers based on single-word utterances during a nine-day training period. Following talker-identification training, listeners were asked to transcribe novel words that were presented in noise; for some listeners, the novel words were produced by the same talkers as used during training and for other listeners the novel words were produced by different talkers than used during training. Transcription scores were higher for the familiar talkers compared to the unfamiliar talkers, indicating that comprehension of spoken words was influenced by previous exposure to particular talkers' voices. The processing benefits of talker familiarity hold when listeners learn to identify talkers on the basis of sentences (Nygaard & Pisoni, 1998) and can be achieved even with short periods of exposure (Bradlow & Pisoni, 1999; Clarke & Garrett, 2004).

Research on perceptual learning in speech suggests that the word recognition benefits associated with talker familiarity might result, at least in part, from adjustments listeners make at a prelexical level of representation (e.g., Norris et al., 2003). Explicit memory tasks have shown that listeners can track a single acoustic-phonetic property on a talker-specific basis (Allen & Miller, 2004), and more implicit phonetic categorization tasks have shown that listeners adjust phonetic boundaries in order to accommodate the idiosyncratic productions of individual talkers (e.g., Eisner & McQueen, 2005). The boundary adjustments associated with perceptual learning occur with minimal exposure (Kraljic &

Samuel, 2007), can persist at least up to a 12-hour delay (Eisner & McQueen, 2006), and influence recognition of novel words (McQueen et al., 2006).[1]

In order to provide a theoretical account of speech perception that describes the encoding and subsequent processing of talker-specific phonetic detail, comprehensive data on the acoustic-phonetic consequences of talker differences in speech production are necessary. In this paper we examine talker differences for one phonetically relevant property of speech, voice-onset-time (VOT). VOT is a primary cue marking the linguistic contrast of voicing in word-initial English stops. In word-initial position, English voiced stops (/b/, /d/, /g/) are typically produced with short VOTs (or, in some cases, with prevoicing), and English voiceless stops (/p/, /t/, /k/), which are aspirated, are produced with longer VOTs (Lisker & Abramson, 1964). Recent research has shown that this property is subject to individual talker differences (Allen et al., 2003). Focusing on voiceless stops, Allen et al. compared word-initial VOTs for many

[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.

monosyllabic words across eight talkers. Their results showed that even after statistically controlling for contextual factors such as speaking rate (using both syllable duration and, in separate analyses, vowel duration as metrics of speaking rate), a statistically significant amount of variability in VOT was accounted for by stable differences across individual talkers. In other words, the talkers differed in their characteristic VOTs, with some talkers producing longer VOTs compared to other talkers.

In this paper we build on this finding by examining the role of contextual influences on VOT at the level of individual talkers. It is well known that VOT is robustly influenced by context (e.g., Klatt, 1975; Lisker & Abramson, 1967; Picheny et al., 1986; Robb et al., 2005). Two contextual factors that have been examined extensively, and that are the focus of the current research, are speaking rate and place of articulation. With respect to speaking rate, it is well documented that VOT systematically increases as speaking rate slows (and syllables become longer), especially for voiceless aspirated stops such as English /p/, /t/, /k/ (e.g., Kessinger & Blumstein, 1997; Miller et al., 1986; Nagao & de Jong, 2007). What is currently unknown is whether the increase is the same magnitude for all talkers, or whether talkers exhibit systematic variability in the extent to which changes in rate affect VOT. Similarly, with respect to place of articulation, it is well established that, in general, VOT increases as place moves from an anterior to posterior point of constriction in the vocal tract (e.g., Lisker & Abramson, 1964; Cho & Ladefoged, 1999; Volaitis & Miller, 1992). Again, what is currently unknown is whether talkers systematically vary in the magnitude of this effect. We examine these two questions in the current work. To preview our results, we find that the magnitude of the speaking rate effect is talker-specific, whereas the magnitude of the place effect is not. The implications of these distinct patterns of results for accounts of speech perception are considered in the Discussion section of the paper.

We report two experiments. Experiment 1 is centered on the effect of speaking rate on VOT in the context of the alveolar voiceless stop /t/. In Experiment 2, we extend the findings of Experiment 1 to the labial (/p/) and velar (/k/) voiceless stops, as well as examine the effect of place of articulation per se on VOT.

2.2 Experiment 1

The primary goal of Experiment 1 was to extend the investigation of talker differences in word-initial VOT for voiceless stop consonants by examining the effect of speaking rate on VOT at the level of individual talkers. Specifically, we examined whether the magnitude of the increase in VOT as speaking rate slows systematically differs across talkers. A secondary goal of Experiment 1 was to replicate Allen et al. (2003) using a different methodology. They observed talker differences in VOT when speaking rate was statistically controlled; we examined whether such differences are also observed when comparing syllables produced at the same rate of speech.[2]

2.2.1 Methods

Subjects

Ten talkers (5 male, E1M1 – E1M5; 5 female, E1F1 – E1F5) were recruited from the Northeastern University community for this experiment. The talkers were native speakers of American English between 18 and 31 years of age with no history of speech or language disorders, and were either paid or received partial course credit for their participation.

Recordings

A magnitude-production procedure (e.g., Adams et al., 1993; Lane & Grosjean, 1973; Miller et al., 1986; Volaitis & Miller, 1992) was used to elicit multiple repetitions of the syllable /ti/ that span a wide range of syllable durations. The alveolar stop was recorded in a constrained phonetic environment in order to control for factors that can influence VOT and syllable duration (e.g.,

[2] Speaking rate, at a global level, is a complex variable. It encompasses not only the rate at which speech itself is produced, but also the number and duration of pauses, as well as aspects of higher-level prosodic structure. There is evidence that the specific way in which a change in speaking rate is implemented may vary in numerous respects across individual talkers (e.g., Crystal & House, 1982; Kuehn & Moll, 1976; McClean, 2000; Matthies et al., 2001). Nonetheless, it appears that for all talkers a change in overall rate involves a change in the rate of speech itself (Miller et al., 1984), however it is implemented at an articulatory level, and this is the focus of the current study. Specifically, we examine how the rate at which a syllable was produced (defined in terms of its syllable or vowel duration, see main text) influences VOT.

vowel identity and final consonant, Port & Rotunno, 1979; Weismer, 1979); such factors could introduce extraneous variability, making it difficult to isolate talker-specific effects of rate on VOT. In the magnitude-production procedure, talkers were directed to produce clear tokens of the syllable /ti/ at their normal speaking rate and at rates relative to their normal speaking rate. Each talker was recorded producing eight runs of syllables. A run consisted of six repetitions of /ti/ at each of the following speaking rates: normal, twice as fast, four times as fast, as fast as possible, normal, twice as slow, four times as slow, as slow as possible. Thus, each run yielded syllables produced at eight speaking rates – seven unique speaking rates and two blocks of repetitions produced at a normal speaking rate. Note that this procedure was used as a tool for acquiring syllables that exhibited variation in overall duration and not as a means to compare the duration of syllables across individuals produced, for example, at a normal speaking rate. The extreme rate prompts (e.g., as fast as possible) were provided to encourage duration variation, and talkers were told that these prompts should reflect the variation found in natural speech and not, for example, the direction to speak as fast as humanly possible. Talkers were given a practice run prior to the recording session and were also given a short break after the first four runs. All recordings took place in a sound-attenuated booth. Speech was recorded via microphone (AKG C460B) onto digital audiotape.

In total, 3840 syllables (6 repetitions X 8 speaking rates X 8 runs X 10 talkers) were recorded. All recordings were digitized at a sampling rate of 20 kHz using the CSL system (KayPENTAX). Syllables produced in the first block of the normal speaking rate for each run were excluded from further analyses to help ensure that, at least to a first approximation, tokens were evenly distributed across the measured range of syllable (and vowel) duration. In addition, the final repetition at each speaking rate was excluded from further analyses because this token may have been subject to a phrase-final lengthening effect (Klatt, 1976). Excluding these tokens left 2800 possible syllables (5 repetitions X 7 speaking rates X 8 runs X 10 talkers) for acoustic analysis.

Acoustic measurements

The Praat speech analysis software (Boersma, 2001) was used to generate a waveform for each syllable. On each waveform, three points in time were located: the onset of the release burst, marked by the onset of low amplitude, aperiodic noise; voicing onset, marked by the onset of high amplitude, periodic energy; and voicing offset, marked by the offset of the last visible glottal pulse. From these three points in time, three durations were calculated. VOT was calculated as the latency between the release burst and voicing onset. Vowel duration was calculated as the latency between voicing onset and voicing offset. Syllable duration was calculated as the latency between the onset of the release burst and voicing offset. In line with numerous studies examining the effect of speaking rate at the segmental level, vowel duration and syllable duration were used as metrics of rate (e.g., Allen et al., 2003; Kessinger & Blumstein, 1997; Nagao & de Jong, 2007; Port, 1981). Vowel duration was used as the primary metric because the statistical analyses used in the current research require that the metric of speaking rate and VOT be mathematically independent. (Because the syllable duration measurement for a particular token includes VOT for that token, syllable duration is not mathematically independent of VOT.) However, syllable duration was also considered, as a secondary metric, in accord with the traditional definition of speaking rate as number of syllables produced per unit time. For all analyses presented in this paper, two versions were conducted; one that used vowel duration as the metric of rate and one that used syllable duration as the metric of rate. Analogous results were found in all cases. For ease of explication, we describe all analyses and results only using the vowel duration metric.

For the 2800 syllables measured, two exclusionary criteria were used to select a final set for statistical analysis. First, a token was excluded if there were production anomalies or if a clear burst onset and vowel offset could not be determined; 2.4% of the tokens were excluded on this basis. Second, a token was excluded if its syllable duration was greater than 799 ms. This criterion, which was established through informal listening, was intended to exclude tokens that were perceived as unnaturally long; 2.8% of the tokens were excluded on this basis, yielding 2654 syllables that spanned durations from 125 ms to 798 ms for use in subsequent analyses.

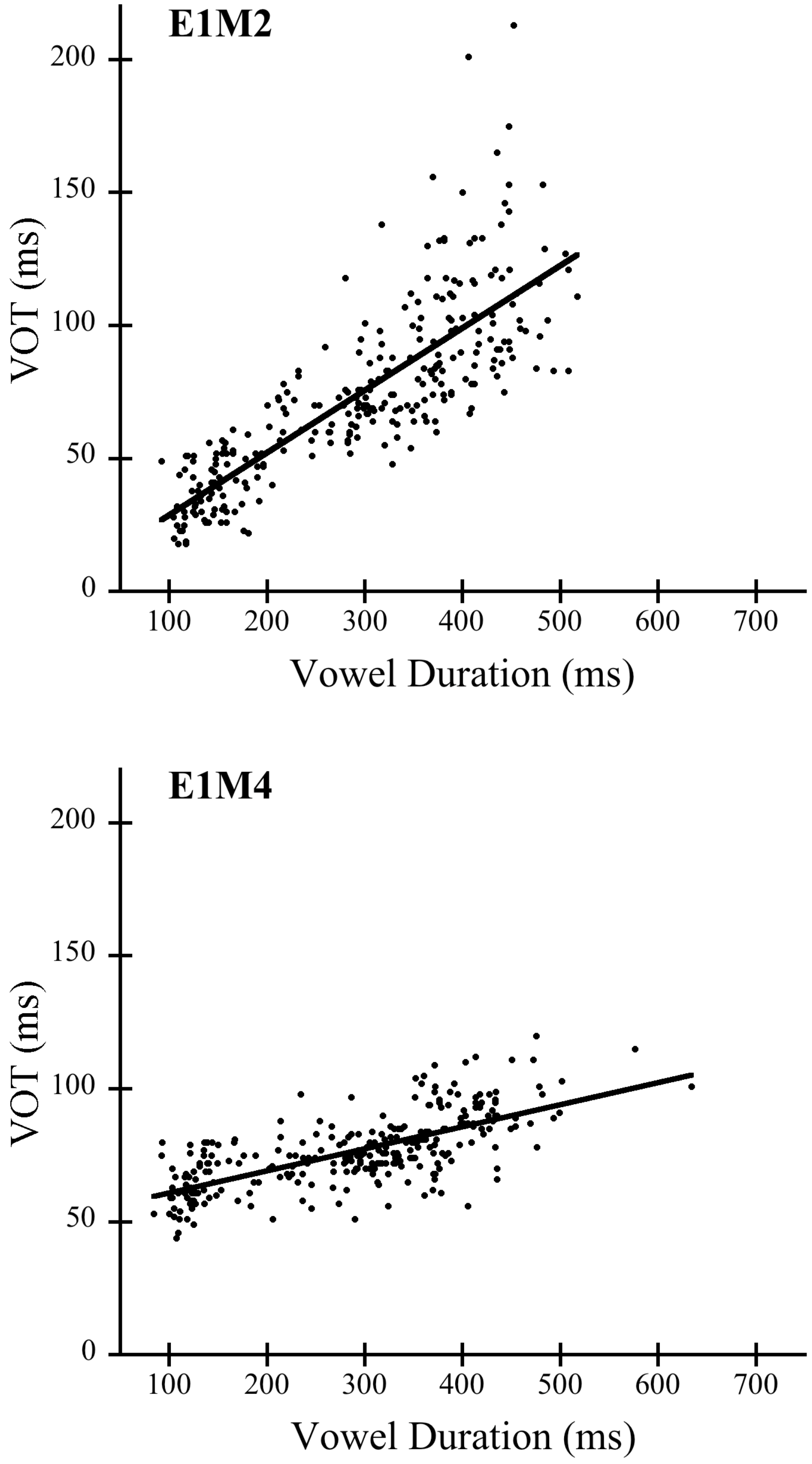
Reliability

One trained experimenter conducted all acoustic measurements. In order to determine cross-experimenter reliability, a different trained experimenter measured approximately 13% of the syllables (one randomly determined run from each talker). Correlations (Pearson's *r*) between the two experimenters' measurements were 0.99 for both VOT and vowel duration. The mean absolute difference between the experimenters' measurements was 2 ms (SD = 2) for VOT and 12 ms (SD = 16) for vowel duration.

2.2.2 Results

For each of the 10 talkers, a linear function relating VOT to vowel duration was calculated using a least squares prediction method. To illustrate, Figure 2.1 shows VOT (ms) as a function of vowel duration (ms) for two of the 10 talkers; in this figure, each filled circle represents a single token of /ti/ and the solid lines represent the linear functions relating VOT to vowel duration. For

Figure 2.1: VOT (ms) as a function of vowel duration (ms) for talkers E1M2 (top panel) and E1M4 (bottom panel). In both panels, each filled circle represents one token of /ti/ and the solid line represents the linear function relating VOT to vowel duration.



both talkers, the tokens span a wide range of vowel durations, and VOT systematically increases as speaking rate slows.[3]

Table 2.1 shows the slopes and intercepts of the 10 individual talker functions, as well as the correlations (Pearson's *r*) between the functions and observed values as an index of goodness-of-fit. Slopes are shown as the increase in VOT (ms) per 100 ms increase in vowel duration and the intercepts are shown as VOT at the mean vowel duration produced across all talkers, which was 319 ms. The slopes of the individual talker functions measure the effect of speaking rate on VOT. The intercepts of the individual talker functions represent VOT at a given vowel duration; in other words, the intercepts of the individual talker functions measure VOT at a single speaking rate.

Consider first the slopes of the individual talker functions. Across the 10 talkers, the slopes show wide variability. For example, given a 100 ms change in vowel duration, VOT for talker E1M2 increases approximately three times as

[3] As described in the main text, one assumption of the statistical analyses used in the current research is that VOT and the metric of speaking rate (e.g., vowel duration) are mathematically independent. An additional assumption is that the relationship between VOT and the metric of speaking rate can be adequately described as linear. For the range of speaking rates that occur in typical speech, there is no established theoretical relationship between VOT and speaking rate. To ensure that a linear function would adequately describe the relationship between VOT and speaking rate for each of the 10 talkers in the current study, we compared three different functions using both vowel duration and syllable duration as the metric of speaking rate: a linear function, an exponential function (with VOT on a linear scale and speaking rate on a log scale), and a power function (with both VOT and speaking rate on a log scale). In all 20 cases (10 talkers X 2 metrics of speaking rate), the correlation coefficient (Pearson's *r*) associated with the linear function was statistically significant, and, critically, was greater than or statistically equal to the correlation coefficient of the exponential and power functions.

much as VOT for talker E1M4 (also shown in Figure 2.1). Turning to the intercepts of the individual talker functions, VOT also varies considerably, spanning values from 62 ms to 91 ms. Inspection of these parameters suggests that the magnitude of the effect of rate on VOT does vary across talkers, and that talker differences in VOT are present for syllables produced at the same speaking rate.

Table 2.1: Slope, intercept, and correlation (Pearson's *r*) of the alveolar functions for individual talkers. Slopes are shows as VOT (ms) / 100 ms vowel duration. The intercepts reflect VOT (ms) at 319 ms vowel duration. Experiment 1.

	Alveolar		
Talker	Slope	Intercept	r
E1M1	21	91	0.78
E1M2	23	79	0.81
E1M3	14	62	0.67
E1M4	8	78	0.69
E1M5	7	62	0.50
E1F1	16	77	0.68
E1F2	10	82	0.71
E1F3	22	86	0.70
E1F4	14	71	0.71
E1F5	12	87	0.74

An HLM analysis (Bryk & Raudenbush, 1992) was used in order to test the statistical significance of the variability in talkers' slopes and intercepts. One benefit of using an HLM analysis is that it allows us to compare the slope and intercept parameters across talkers while taking into account the entire set of data, which consisted of 2654 tokens. (A complete description of the HLM structure for all models presented in this paper is provided in the Appendix.) In terms of the talkers' slopes, results showed that the mean slope across talkers was non-zero [$t(9) = 8.11$, $p < .001$], which confirms that, as expected, VOT systematically increased as vowel duration increased (i.e., rate slowed) across the group of talkers. Critically, the results also showed that there was significant variability in the talkers' slopes [$\chi^2(9) = 374.78$, $p < .001$], indicating that how much VOT increased as rate slowed was not the same for all talkers. In terms of the talkers' intercepts, results confirmed that the mean intercept across talkers was non-zero [$t(9) = 25.43$, $p < .001$], as expected, and that there was significant variability in the talkers' intercepts [$\chi^2(9) = 776.23$, $p < .001$]. This finding indicates that talkers differed in their characteristic VOTs for utterances produced at the same speaking rate.

An additional set of analyses was performed in order to examine whether talker differences in VOT would be observed across a range of vowel durations, and not solely at the mean vowel duration produced across all talkers. The motivation for these analyses stems from the finding that there was significant variability in the slopes of the individual talker functions, with some functions intersecting within the measured range of vowel duration. As a consequence, even though talker differences in VOT were observed at the mean vowel duration, they will not necessarily be observed across a range of vowel durations. For these analyses, four intercepts (shown in Table 2.2) were calculated for each talker corresponding to VOT (ms) at 200, 300, 400, and 500 ms vowel duration; these values span the range of greatest intersection among the individual functions.

Table 2.2: Intercepts of the alveolar functions for individual talkers, defined as VOT (ms) at 200, 300, 400, and 500 ms vowel duration. Experiment 1.

	Alveolar Intercepts			
	Vowel Duration			
Talker	200	300	400	500
E1M1	66	87	108	129
E1M2	51	74	97	120
E1M3	45	59	73	87
E1M4	69	77	85	93
E1M5	54	61	68	75
E1F1	58	74	90	106
E1F2	71	81	91	101
E1F3	60	82	104	126
E1F4	54	68	82	96
E1F5	73	85	97	109

HLM analyses (see Appendix) confirmed that there was significant variability in talkers' intercepts at each vowel duration [in all cases; $\chi^2(9) > 311.00$, $p < .001$], indicating that the presence of talker differences in VOT is not contingent on speaking rate.

2.3 Experiment 2

The results from Experiment 1, which focused on the production of word-initial /t/, confirm that at a given rate of speech, talkers differ in their characteristic VOTs. They also confirm that speaking rate influences VOT, such that as speaking rate slows VOT increases. With regard to the primary focus of Experiment 1, the results also indicate that the contextual effect of speaking rate on VOT is itself talker-specific; the magnitude of the increase in VOT as speaking rate slows varies across individual talkers. In Experiment 2, we use the same basic procedures used in Experiment 1 to extend these findings in three ways.

First, we attempt to replicate the findings from Experiment 1 for the other two voiceless stops in English, labial /p/ and velar /k/. Second, we examine whether the effect of rate for a particular talker is stable across a change in place of articulation by comparing the slopes of the functions relating VOT to vowel duration for the labial and velar voiceless stops. Third, Experiment 2 examines whether the contextual influence of place of articulation on VOT is itself talker-specific. As noted earlier, previous research has shown that, in general, VOT increases as place moves from front to back in the vocal tract; for example, VOT for /k/ is typically longer than VOT for /p/ (e.g., Lisker & Abramson, 1964). In the current experiment we examine whether this effect, like the effect of speaking rate examined in Experiment 1, is talker-specific. That is, we examine whether the magnitude of the difference in VOT between /p/ and /k/ varies across individual talkers.

2.3.1 Method

Subjects

Ten talkers (5 male, E2M1 – E2M5; 5 female, E2F1 – E2F5) who did not participate in Experiment 1 were recruited from the Northeastern University community for this experiment. The talkers were native speakers of American English between 18 and 22 years of age with no history of speech or language disorders, and were either paid or received partial course credit for their participation.

Recordings

The magnitude-production procedure described in Experiment 1 was used to elicit multiple repetitions of the syllables /pi/ and /ki/ across a range of syllable durations. As in Experiment 1, talkers produced eight runs of each syllable, with each run consisting of six repetitions at eight speaking rates. The order of the labial and velar syllables was counter-balanced across talkers. All recordings followed the procedure outlined for Experiment 1.

In total, 7680 syllables (6 repetitions X 8 speaking rates X 8 runs X 10 talkers X 2 places of articulation) were recorded. All recordings were digitized at a sampling rate of 20 kHz using the CSL system. As in Experiment 1, all syllables produced in the first block of the normal speaking rate for each run and the final repetition at each speaking rate were excluded from further analyses. Excluding these tokens left 5600 possible syllables (5 repetitions X 7 speaking rates X 8 runs X 10 talkers X 2 places of articulation) for acoustic analysis.

Acoustic measurements

The Praat speech analysis software was used to generate a waveform for each of the 5600 syllables. As in Experiment 1, VOT, vowel duration, and syllable duration were calculated for each waveform, and two exclusionary criteria were used to select a final set of syllables for statistical analysis. First, a token was excluded if there were production anomalies or if a clear burst onset and vowel offset could not be determined; 7.4% of the tokens were excluded on this basis. Second, a token was excluded if its syllable duration was greater than 799 ms; 9.6% of the tokens were excluded on this basis. As a result of this selection process, 4646 syllables that spanned durations from 115 ms to 799 ms were used in subsequent analyses.[4]

Reliability

Two trained experimenters, who each measured a subset of the recorded tokens, conducted all acoustic measurements. To determine cross-experimenter reliability, a third trained experimenter measured one randomly determined run

[4] As is apparent, a larger percentage of tokens was excluded from statistical analysis in Experiment 2 than Experiment 1, due both to an increased proportion of anomalous/unmeasurable tokens and to an increased proportion of extremely long tokens. The underlying reason for the difference across experiments is not known. Importantly, even with the exclusion, the number of tokens available for statistical analysis in both experiments was very large.

of /pi/ and /ki/ for each talker (approximately 13% of the tokens). Correlations (Pearson's r) between the two experimenters' measurements were 0.98 for VOT and 0.99 for vowel duration. The mean absolute difference between the experimenters' measurements was 4 ms (SD = 6) for VOT and 29 ms (SD = 27) for vowel duration.

2.3.2 Results

For each of the 10 talkers, two linear functions relating VOT to vowel duration were calculated using a least squares prediction method, one for the labial syllables and one for the velar syllables.[5] Table 2.3 shows the slopes and intercepts of the individual talker functions, as well as the correlations (Pearson's r) between the functions and observed values as an index of goodness-of-fit. Slopes are shown as the increase in VOT (ms) per 100 ms increase in vowel duration and the intercepts are shown as VOT at the mean vowel duration produced across all talkers for the labial and velar tokens, which was 374 ms. Three sets of analyses were performed on the parameters specifying the individual talker functions. In the first set of analyses, we attempted to extend

[5] As in Experiment 1, we confirmed that the relationship between VOT and the metric of speaking rate could be adequately described as linear for the 10 talkers examined here. For each place of articulation, we examined the correlation coefficient (Pearson's r) of three different functions (linear, exponential, power) using both vowel duration and syllable duration as the metric of speaking rate. In all 40 cases (10 talkers X 2 places of articulation X 2 metrics of speaking rate), the correlation coefficient of the linear function was statistically significant and, critically, was better than or statistically equal to the correlation coefficient of the exponential and power functions.

findings from Experiment 1 to labial and velar voiceless stops. The second set of analyses examined whether, for a given talker, the magnitude of the effect of speaking rate on VOT is stable across place of articulation. The third set of analyses examined whether the contextual influence of place of articulation itself is talker-specific.

Table 2.3: Slope, intercept, and correlation (Pearson's r) of the labial and velar functions for individual talkers. Slopes are shown as VOT (ms) / 100 ms vowel duration. The intercepts reflect VOT (ms) at 374 ms vowel duration. Experiment 2.

	Labial			Velar		
Talker	Slope	Intercept	r	Slope	Intercept	r
E2M1	9	60	0.55	6	103	0.32
E2M2	5	37	0.48	10	91	0.55
E2M3	25	83	0.80	20	99	0.81
E2M4	10	55	0.74	7	95	0.53
E2M5	3	30	0.42	4	86	0.48
E2F1	10	78	0.51	13	111	0.44
E2F2	10	64	0.77	9	93	0.71
E2F3	12	77	0.48	13	126	0.61
E2F4	3	57	0.30	6	81	0.58
E2F5	16	81	0.41	23	113	0.59

Replication

The first set of analyses examined whether, for a given voiceless stop, talkers differed in their characteristic VOTs at a single speaking rate and, critically, whether the magnitude of the effect of rate on VOT varied across talkers. Thus, two analyses were performed, one for the labial functions and one for the velar functions. HLM analyses were applied to the labial data (2481 tokens) and, separately, to the velar data (2165 tokens), following the structure used for Experiment 1. Results showed that for both the labial and velar functions, there was significant variability in the slopes [$\chi^2(9) = 481.47$, $p < .001$ and $\chi^2(9) = 332.37$, $p < .001$; respectively] and intercepts [$\chi^2(9) = 1559.82$, $p < .001$ and $\chi^2(9) = 957.19$, $p < .001$; respectively] of the individual talker functions. These results extend the findings from Experiment 1 to labial and velar voiceless stops, confirming not only the presence of talker differences in VOT at a single speaking rate, but also that the effect of speaking rate on VOT varied significantly across talkers. As in Experiment 1, we also tested for talker differences in VOT across a range of vowel durations (i.e., speaking rates) for both the labial and velar stops (see Table 2.4). HLM analyses showed that there was significant variability in talkers' intercepts at each vowel duration [in all cases; $\chi^2(9) > 274.00$, $p < .001$].

Stability of the effect of speaking rate on VOT for individual talkers

Results reported above indicate that the magnitude of the effect of speaking rate on VOT varies across talkers for a given voiceless stop. This finding highlights a source of systematic variability in the speech signal in that how much VOT increases as rate slows can vary from talker to talker. Here we examine a potential source of stability in the speech signal by comparing the effect of rate on VOT for a given talker across a change in place of articulation.

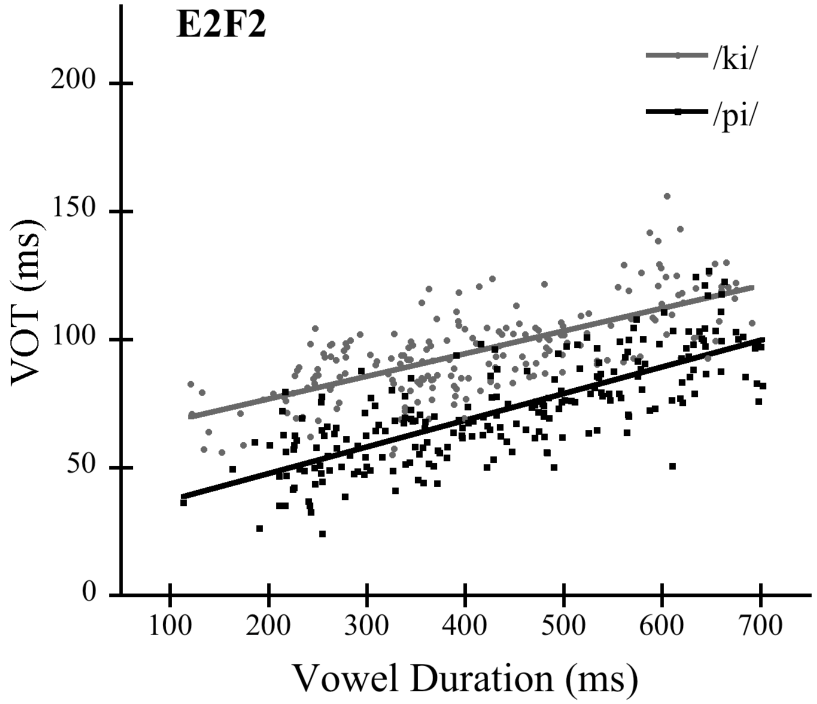
Table 2.4: Intercepts of the labial and velar functions for individual talkers, defined as VOT (ms) at 200, 300, 400, and 500 ms vowel duration. Experiment 2.

	Labial Intercepts				Velar Intercepts			
	Vowel Duration				Vowel Duration			
Talker	200	300	400	500	200	300	400	500
E2M1	44	53	62	71	93	99	105	111
E2M2	28	33	38	43	74	84	94	104
E2M3	40	65	90	115	64	84	104	124
E2M4	37	47	57	67	83	90	97	104
E2M5	24	27	30	33	79	83	87	91
E2F1	60	70	80	90	88	101	114	127
E2F2	47	57	67	77	77	86	95	104
E2F3	57	69	81	93	103	116	129	142
E2F4	51	54	57	60	71	77	83	89
E2F5	53	69	85	101	73	96	119	142

In this analysis, we considered the slopes of the labial and velar functions for individual talkers. If, for a given talker, the effect of rate on VOT is stable across a change in place of articulation, then the slopes of the labial and velar functions will be approximately the same. Inspection of the labial and velar slopes, shown in Table 2.3, suggests this may be the case in that the difference between the labial and velar slopes for any given talker is quite small. To illustrate, Figure 2.2 displays VOT (ms) as a function of vowel duration (ms) at both places of articulation for one of the 10 talkers. VOT increases as speaking rate slows for both the labial and velar tokens, and does so to approximately the same degree.

In order to examine the statistical significance of the difference between the labial and velar slopes for individual talkers, two analyses were performed. First, a paired t-test revealed that the mean difference between the labial and velar slopes (0.80) was non-significant [$t(9) = 0.66$, $p = .52$]. This analysis indicates that there was no systematic effect of place of articulation on the slopes for the group of talkers. However, this analysis is potentially misleading in terms of describing whether, for a given talker, the slopes of the labial and velar functions were the same. The non-significant mean difference for the group of talkers could indicate that for some talkers the labial slope was greater than the velar slope and for other talkers the labial slope was less than the velar slope, rather than indicating that the labial and velar slopes were the same for a given talker. In order to ensure that this was not the case, we conducted an additional HLM analysis nesting the labial and velar slopes within talkers. Results from this analysis revealed that there was no significant variability across talkers in the difference between the labial and velar slopes [$\chi^2(9) = 1.60$, $p > .50$], which indicates that the effect of speaking rate on VOT for a given talker is the same for labial and velar voiceless stops. In other words, the effect of rate on VOT

Figure 2.2: VOT (ms) as a function of vowel duration (ms) for talker E2F2 at two places of articulation, labial /pi/ and velar /ki/. Each black square represents one token of /pi/ and each grey circle represents one token of /ki/. The solid lines represent the calculated function relating VOT to vowel duration at each place of articulation.



varies across talkers within a given place of articulation, but the effect of rate on VOT for a particular talker is stable in that it holds across a change in place of articulation.

Effect of place of articulation on VOT for individual talkers

The third set of analyses addressed whether the effect of place of articulation on VOT itself varies across individual talkers. Previous research has shown that VOTs for labial stops are shorter than VOTs for velar stops (e.g., Lisker & Abramson, 1964). Here we examined whether the magnitude of the difference between labial and velar VOTs varies significantly across talkers. To quantify the effect of place of articulation on VOT for each talker, we used the difference between the labial and velar intercepts, with the intercept defined as VOT at 374 ms vowel duration (shown in Table 2.3). Because the results reported above indicate that the slopes of the labial and velar functions within a given talker are not statistically different (and thus the functions are approximately parallel), using a single point on each function as the basis of comparison is valid in that the difference between labial and velar VOTs will be the same for any value along the x-axis. As expected, the labial intercept was located at a shorter VOT than the velar intercept for each talker, resulting in a reliable effect of place of articulation on VOT for the group of talkers [mean difference = 37.60 ms; $t(9) = 9.06$, $p < .001$]. To examine the central question of whether the magnitude of the difference in labial and velar intercepts varied significantly across talkers, an HLM analysis was used to nest the labial and velar intercepts within talkers. The HLM results showed that there was no significant variability in the difference between the labial and velar intercepts across individual talkers [$\chi^2(9) = 2.97$, $p > .50$].

These results indicate that the effect of place of articulation on VOT does not vary across individual talkers; rather, the magnitude of displacement between labial and velar VOTs was approximately the same for each talker.

2.4 Discussion

Previous research has provided evidence for talker-specific variability in the acoustic-phonetic information used to convey individual speech segments (e.g., Espy-Wilson et al., 2000; Newman et al., 2001; Peterson & Barney, 1952; Zue & Laferriere, 1979). As a case in point, recent findings indicate that talkers differ in VOTs produced for voiceless stop consonants; some talkers produce characteristically shorter VOTs than other talkers (Allen et al., 2003). The results from the current research, which examined /ti/ in Experiment 1 and /pi/ and /ki/ in Experiment 2, confirm this finding and, most importantly, extend it by examining potential talker specificity in how two contextual variables, speaking rate and place of articulation, influence VOT.

In terms of speaking rate, previous research has shown that as speaking rate slows (and syllables become longer), VOT systematically increases (e.g., Kessinger & Blumstein, 1997; Volaitis & Miller, 1992). The current results replicate this finding for all three voiceless stops. However, the results also showed that for each stop, the magnitude of the increase in VOT for a given change in speaking rate varied significantly across talkers. This finding, which indicates that the effect of speaking rate on VOT is talker-specific, highlights a source of systematic variability in the speech signal. Further, the results from Experiment 2, which compared /p/ and /k/, showed that for a given talker, the magnitude of the rate effect on VOT remained constant across a change in place of articulation for the CV syllables examined in the current experiments. This finding highlights a source of stability in the speech signal at the individual talker level, in that how rate influences VOT for one voiceless stop is the same for a different voiceless stop.

In terms of place of articulation, previous research has shown that VOT systematically increases as place moves from an anterior to posterior constriction in the vocal tract (e.g., Lisker & Abramson, 1964). In line with this finding, the results from Experiment 2 showed that for each talker VOTs for /p/ were shorter than VOTs for /k/. Critically, the results also indicated that the magnitude of displacement between VOTs for /p/ and /k/ did not vary significantly across talkers in the case examined in the current experiments. Thus, unlike speaking rate, the contextual influence of place of articulation on VOT appears not to be talker-specific.

These findings have implications for theoretical accounts of speech perception. As reviewed in the Introduction, there is evidence that listeners retain talker-specific acoustic-phonetic information in memory (e.g., Goldinger, 1998) and that familiarity with a particular talker's speech can facilitate word recognition (e.g., Nygaard et al., 1994). Furthermore, findings from the literature on perceptual learning in speech suggest that the benefits of talker familiarity observed at the word level might result, at least in part, from talker-specific effects at a prelexical level of representation (e.g., Eisner & McQueen, 2005; Kraljic & Samuel, 2007). Of particular relevance to the current research, Allen and Miller (2004) showed that listeners could learn that one talker produces a particular voiceless stop with characteristically short VOTs and a different talker produces the same stop with characteristically long VOTs.

This finding raises the possibility that listeners may customize stop voicing categories based on individual talkers' characteristic VOTs. However, given contextual influences on VOT, listeners would need to consider a talker's characteristic VOTs not in an absolute manner, but with respect to context. Indeed, it is well established that at a general level, listeners do process VOT in relation to numerous contextual factors, including both speaking rate and place of articulation (e.g., Lisker & Abramson, 1970; Miller & Volaitis, 1989; Summerfield, 1981; Volaitis & Miller, 1992). We do not yet know whether such context-dependent processing is tuned to the speech of individual talkers, but the results of the current experiments place constraints on the type of exposure listeners might require for such perceptual tuning.

Specifically, the current data suggest that exposure to a talker's VOTs for a voiceless stop at one speaking rate would not optimally inform the listener as to that talker's VOTs for the stop at a novel speaking rate. Because the magnitude of the rate effect systematically varies across talkers, in order to accommodate the contextual influence of rate on VOT listeners would need to learn, for a given talker, how much VOT changes as a function of speaking rate; that is, ascertain the slope of the function relating VOT to rate. However, because the magnitude of the rate effect on VOT for a given talker is stable across a change in place of articulation, tracking the contextual influence of rate in the context of one voiceless stop could potentially inform the listener as to how this contextual influence operates for other voiceless stops in similar phonetic environments.

The current data also suggest that listeners might not need to track the contextual influence of place of articulation per se on VOT at the level of individual talkers. Because the magnitude of the place effect does not systematically differ across individual talkers, listeners could rely on more general knowledge, perhaps specific to their language (e.g., Cho & Ladefoged, 1999), to inform them as to how VOT shifts as a function of place of articulation. As a consequence, for a given speaking rate and a similar phonetic environment, learning a particular talker's characteristic VOTs for one voiceless stop could potentially inform the listener as to that talker's VOTs for voiceless stops with a different place of articulation.

In sum, the present data provide basic information on how two contextual factors influence VOT at a talker-specific level and, in so doing, point to constraints on how listeners might accommodate such contextual variation when customizing phonetic categories for an individual talker's speech. Future research is aimed at examining the nature and extent of such perceptual fine-tuning.

Appendix to Chapter 2

HLM analyses (Bryk & Raudenbush, 1992) were used in the current research because they allow examination of stable individual differences around group level patterns. HLM analyses are based on linear regression techniques; however, unlike standard regression models, HLM analyses are well suited for examination of data from repeated-measures designs. All of the analyses reported in this chapter are based on two HLM structures. The first HLM structure was used to compare slope and intercept parameters within a single place of articulation. This model was used in Experiment 1 to compare the slopes and intercepts of the alveolar functions and in Experiment 2 to compare the slopes and intercepts of the labial functions and, separately, the velar functions. The second HLM structure was used to compare the slope and intercept parameters across place of articulation. This structure was used in Experiment 2 to compare the slopes of the labial functions to the slopes of the velar functions, and, separately, to compare the intercepts of the labial functions to the intercepts of the velar functions. The details of each type of model are presented in turn.

In order to test the statistical significance of the variability in talkers' slopes and intercepts within a single place of articulation (e.g., /ti/), all of the tokens for the particular analysis were nested within each of the 10 talkers as follows:

Level 1 model:

$$\mathrm{VOT}_{ij} = \beta_{0j} + \beta_{1j}\ (\text{vowel duration}) + r_{ij}$$

Level 2 model:

$$\beta_{0j} = \gamma_{00} + u_{0j}$$

$$\beta_{1j} = \gamma_{10} + u_{1j}$$

With this structure, VOT is specified as a function of vowel duration, while incorporating the fact that sets of individual tokens are associated with specific talkers. Importantly, the Level-2 model allows the intercepts (β_{0j}) and slopes (β_{1j}) of the Level-1 model to vary across talkers. That is, the Level-2 model estimates the mean intercept (γ_{00}) and mean slope (γ_{10}) values across talkers while also testing if significant variability exists in these parameters (u_{0j} and u_{1j}, respectively) as a function of stable talker differences.

In order to examine the slopes (or intercepts) across place of articulation, the labial and velar slopes (or intercepts) were nested within talkers as follows:

Level 1 model:

$$\text{Slope (or intercept)}_{ij} = \beta_{0j} + \beta_{1j}\ (\text{place of articulation}) + r_{ij}$$

Level 2 model:

$$\beta_{0j} = \gamma_{00}$$

$$\beta_{1j} = \gamma_{10} + u_{1j}$$

With this structure, slope (or intercept) is specified as a function of place of articulation, while incorporating the fact that pairs of individual values are associated with specific talkers. In order to allow place of articulation to be examined as a linear variable, labial and velar were coded as 0 and 1, respectively. Using this method, the slope parameter of the HLM (β_{1j}) does not indicate the absolute slope (or intercept) for either the labial or velar functions; rather, it represents the difference between the labial and velar slopes (or intercepts). The Level-2 model allows the slope (β_{1j}) of the Level-1 model to vary across talkers; accordingly, the model estimates the mean difference between the labial and velar slopes (or intercepts) across talkers (γ_{10}) while also testing if significant variability exists in this parameter (u_{1j}).

REFERENCES

Adams, S. G., Weismer, G., and Kent, R. D. (1993). Speaking rate and speech movement velocity profiles. *Journal of Speech, Language, and Hearing Research, 36,* 41-54.

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. *Perception & Psychophysics, 63,* 798-810.

Allen, J. S., and Miller, J. L. (2004). Listener sensitivity to individual talker differences in voice-onset-time. *Journal of the Acoustical Society of America, 115,* 3171-3183.

Allen, J. S., Miller, J. L., and DeSteno, D. (2003). Individual talker differences in voice-onset-time. *Journal of the Acoustical Society of America, 113,* 544-552.

Boersma, Paul. (2001). Praat, a system for doing phonetics by computer. *Glot International, 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. *Journal of the Acoustical Society of America, 106,* 2074-2085.

Bryk, A. S., and Raudenbush, S.W. (1992). Hierarchical Linear Models: Applications and Data Analysis Methods (Sage, Newbury Park, CA).

Byrd, D. (1992). Preliminary results on speaker-dependent variation in the TIMIT database. *Journal of the Acoustical Society of America, 92,* 593-596.

Cho, T., and Ladefoged, P. (1999). Variations and universals in VOT: Evidence from 18 languages. *Journal of 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. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 521-533.

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

Crystal, T. H., and House, A. S. (1982). Segmental durations in connected speech signals: Preliminary results. *Journal of the Acoustical Society of America, 72,* 705-716.

Delattre, P. C., Liberman, A. M., and Cooper, F. S. (1955). Acoustic loci and transitional cues for consonants. *Journal of the Acoustical Society of America, 27*, 769-773.

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

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

Eisner, F., and McQueen, J. M. (2006). Perceptual learning in speech: Stability over time. *Journal of the Acoustical Society of America, 119,* 1950-1953.

Espy-Wilson, C. Y., Boyce, S. E., Jackson, M., Narayanan, S., and Alwan, A. (2000). Acoustic modeling of American English /r/. *Journal of the Acoustical Society of America, 108,* 343-356.

Fellowes, J. M., Remez, R. E., and Rubin, P .E. (1997). Perceiving the sex and identity of a talker without natural vocal timbre. *Perception & Psychophysics, 59,* 839-849.

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

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review, 105,* 251-279.

Hashi, M., Honda, K., and Westbury, J. R. (2003). Time-varying acoustic and articulatory characteristics of American English [ɹ]: A cross-speaker study. *Journal of Phonetics, 31,* 3-22.

Hillenbrand, J., Getty, L. A., Clark, M. J., and Wheeler, K. (1995). Acoustic characteristics of American English vowels. *Journal of the Acoustical Society of America, 97,* 3099-3111.

Kessinger, R. H., and Blumstein, S. E. (1997). Effects of speaking rate on voice-onset time in Thai, French, and English. *Journal of Phonetics, 25,* 143-168.

Klatt, D. H. (1975). Voice onset time, frication, and aspiration in word-initial consonant clusters. *Journal of Speech, Language, and Hearing Research, 18,* 686-706.

Klatt, D. H. (1976). Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. *Journal of the Acoustical Society of America, 59,* 1208-1221.

Klatt, D. H. (1986). The problem of variability in speech recognition and in models of speech perception. In J. S. Perkell and D. H. Klatt (Eds.), *Invariance and variability in speech processes* (pp. 300-319). Hillsdale, NJ: Erlbaum.

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

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

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

Kuehn, D. P, and Moll, K. L. (1976). A cineradiographic study of VC and CV articulatory velocities. *Journal of Phonetics, 4,* 303-320.

Kuhl, P. K. (1991). Human adults and human infants show a "perceptual magnet effect" for the prototypes of speech categories, monkeys do not. *Perception & Psychophysics, 50,* 93-107.

Ladefoged, P., and Broadbent, D. E. (1957). Information conveyed by vowels. *Journal of the Acoustical Society of America, 29,* 98-104.

Landahl, K., and Blumstein, S. E. (1982). Acoustic invariance and the perception of place of articulation: A selective adaptation study. *Journal of the Acoustical Society of America, 71,* 1234-1241.

Lane, H., and Grosjean, F. (1973). Perception of reading rate by speakers and listeners. *Journal of Experimental Psychology, 97,* 141-147.

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

Lisker, L., and Abramson, A. S. (1967). Some effects of context on voice onset time in English stops. *Language and Speech, 10,* 1-28.

Lisker, L., and Abramson, A. S. (1970). The voicing dimension: Some experiments in comparative phonetics, in Proceedings of the Sixth International Congress of Phonetic Sciences (Academia, Prague), pp. 563-567.

Matthies, M., Perrier, P., Perkell, J. S., and Zandipour, M. (2001). Variation in anticipatory coarticulation with changes in clarity and rate. *Journal of Speech, Language, and Hearing Research, 44,* 340-353.

McClaskey, C., Pisoni, D. B., and Carrell, T. (1983). Transfer of training of a new linguistic contrast in voicing. *Perception & Psychophysics, 34,* 323-330.

McClean, M. D. (2000). Patterns of orofacial movement velocity across variations in speech rate. *Journal of Speech, Language, and Hearing Research, 43,* 205-216.

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

Miller, J. L. (1981). Effects of speaking rate on segmental distinctions. In P. D. Eimas and J. L. Miller (Eds.), *Perspectives on the study of speech* (pp. 39-74). Hillsdale, NJ: Erlbaum.

Miller, J. L., and Eimas, P. D. (1976). Studies on the selective tuning of feature detectors for speech. *Journal of Phonetics, 4,* 119-127.

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. *Perception and Psychophysics, 46,* 505-512.

Mullennix, J. W., Pisoni, D. B., and Martin, C. S. (1989). Some effects of talker variability on spoken word recognition. *Journal of the Acoustical Society of America, 85,* 365-378.

Nagao, K., and de Jong, K. (2007). Perceptual rate normalization in naturally produced rate-varied speech. *Journal of the Acoustical Society of America, 121,* 2882-2898.

Newman, R. S., Clouse, S. A., and Burnham, J. L. (2001). The perceptual consequences of within-talker variability in fricative production. *Journal of the Acoustical Society of America, 109,* 1181-1196.

Norris, D., McQueen, J. M., and Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology, 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. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26,* 1228-1244.

Nygaard, L. C., and Pisoni, D. B. (1998). Talker-specific learning in speech perception. *Perception & Psychophysics, 60,* 355-376.

Nygaard, L. C., Sommers, M. S., and Pisoni, D. B. (1994). Speech perception as a talker-contingent process. *Psychological Science, 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. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19,* 309-328.

Peterson, G. E., and Barney, H. L. (1952). Control methods used in a study of the vowels. *Journal of the Acoustical Society of America, 24,* 175-184.

Picheny, M. A., Durlach, N. I., and Braida, L. D. (1986). Speaking clearly for the hard of hearing II: Acoustic characteristics of clear and conversational speech. *Journal of Speech, Language, and Hearing Research, 29,* 434-446.

Port, R. F. (1981). Linguistic timing factors in combination. *Journal of the*

Acoustical Society of America, 69, 262-274.

Port, R. F., and Rotunno, R. (1979). Relation between voice-onset time and vowel duration. *Journal of the Acoustical Society of America, 66,* 654-662.

Remez, R. E., Fellowes, J. M., and Rubin, P. E. (1997). Talker identification based on phonetic information. *Journal of Experimental Psychology: Human Perception and Performance, 23,* 651-666.

Remez, R. E., Rubin, P. E., Pisoni, D. B., and Carrell, T. D. (1981). Speech perception without traditional speech cues. *Science, 212,* 947-950.

Robb, M., Gilbert, H., and Lerman, J. (2005). Influence of gender and environmental setting on voice onset time. *Folia Phoniatrica et Logopaedica, 57,* 125-133.

Samuel, A. G. (1982). Phonetic prototypes. *Perception & Psychophysics, 31,* 307-314.

Schacter, D. L., and Church, B. A. (1992). Auditory priming: Implicit and explicit memory for words and voices. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 915-930.

Studdert-Kennedy, M. (1976). Speech perception, in N. J. Lass (Ed.), Contemporary Issues in Experimental Phonetics, New York: Academic Press, 243-293.

Summerfield, Q. (1981). Articulatory rate and perceptual constancy in phonetic perception. *Journal of Experimental Psychology: Human Perception and Performance, 7,* 1074-1095.

Tremblay, K., Kraus, N., Carrell, T. D., and McGee, T. (1997). Central auditory system plasticity: Generalization to novel stimuli following listening training. *Journal of the Acoustical Society of America, 102,* 3762-3773.

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. *Journal of the Acoustical Society of America, 92,* 723-735.

Weismer, G. (1979). Sensitivity of voice-onset time (VOT) measures to certain segmental features in speech production. *Journal of Phonetics,* 7, 197-204.

Zue, V. W., and Laferriere, M. (1979). Acoustic study of medial /t,d/ in American English. *Journal of the Acoustical Society of America, 66,* 1039-1050.