6-13-2013

Talker-specific Influences on Phonetic Boundaries and Internal Category Structure

Janice Ann Lomibao
University of Connecticut - Storrs, janice.lomibao@uconn.edu

Talker-specific Influences on Phonetic Boundaries and Internal Category Structure

Janice Ann Lomibao

B.A., University of Rochester, 2006

A Thesis

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Master of Arts

At the

University of Connecticut

2013

APPROVAL PAGE

Master of Arts Thesis

Talker-specific Influences on Phonetic Boundaries and Internal Category Structure

Presented by

Janice Ann Lomibao

Major Advisor ____
Rachel M. Theodore, Ph.D.

Associate Advisor ____
Carl A. Coelho, Ph.D.

Associate Advisor ____
Emily B. Myers, Ph.D.

University of Connecticut

2013

Acknowledgments

This research was supported the College of Liberal Arts and Sciences (University of Connecticut) in the form of start-up funds to Dr. Rachel M. Theodore. We thank committee members Dr. Carl Coelho and Dr. Emily B. Myers for their support and useful insight in the development of this project. We also thank Devin Roscillo, Sarah Montanaro, and the other members of the UConn Laboratory for Spoken Language Processing for assistance with data collection and analysis.

Table of Contents

Abstract

A major goal of research in the field of speech perception has been to explain how listeners consistently extract individual speech sounds from the speech stream given that there is a lot of variability in the acoustic-phonetic signal for individual consonants and vowels. That is, there is no one-to-one relationship between a given speech sound and the acoustic information specifying a given speech sound. Variability for individual speech sounds comes from many sources including idiosyncratic differences in pronunciation across individual talkers, which is the focus of the current work. Researchers have shown that one way listeners achieve consistent perception despite talker variability is by encoding talker-specific surface characteristics in memory (Goldinger, 1996) and using this information to facilitate linguistic processing including word recognition (Nygaard et al., 1994; Clarke & Garrett, 2004).

There is some evidence that the benefits of talker familiarity observed at higher levels of linguistic processing (e.g., word recognition) reflect adjustments listeners make even earlier in the processing stream. Specifically, there is some evidence that listeners make talker-specific adjustments when recovering the individual phonetic segments, and thus prior to recognizing words. Research in this vein has shown that listeners are sensitive to individual properties of speech on a talker-by-talker basis, including voice-onset-time (VOT) for word-initial stop consonants (Theodore & Miller, 2010). Moreover, listeners will adjust boundaries between phonetic categories in order to accommodate a talker's unique way of producing a phonetic category, at least if the idiosyncratic production is ambiguous between two categories (e.g., Eisner & McQueen, 2005). However, other research, not at the level of individual talkers, has shown that phonetic categories are not marked solely by boundaries. They also have a graded internal structure such that not all members of a phonetic category are considered equally good members. Critically, systematic variation in the speech signal (e.g., speaking rate) strongly influences which members of a phonetic category are considered most prototypical (e.g., Miller & Volaitis, 1989). This finding raises the possibility that listeners may not only shift phonetic boundaries to accommodate a talker's unique productions; they may also shift the internal category structure.

The current work tests the hypothesis that listeners accommodate talker-specific phonetic detail by shifting phonetic category boundaries and internal category structure in line with a talker's characteristic productions. All listeners heard two talkers, Joanne and Sheila, produce the voiceless stop /k/. Listeners were divided into two training groups. One group heard Joanne produce /k/ with short VOTs and Sheila produce /k/ with relatively longer VOTs. The other group heard the opposite pattern of VOT exposure; Joanne produced /k/ with long VOTs and Sheila produced /k/ with relatively shorter VOTs. Exposure to the two talkers occurred during training phases. Following training, listeners were tested using Joanne's voice on three different tasks: (1) a two-alternative forced choice task, in which listeners were presented with a short- and a long-VOT variant and asked to choose which was most representative of Joanne, (2) an identification task, in which listeners were presented with a VOT continuum from *gain* to *cane* and asked to categorize each token as beginning with a voiced or voiceless stop, and (3) a goodness rating task, where listeners were presented with the same VOT continuum used for the identification task and were asked to rate each token for goodness as Joanne’s /k/.

The results showed that listeners were sensitive to talker differences in VOT in that performance during the two-alternative forced choice test was in line with previous training with Joanne’s speech. This pattern is as predicted based on earlier research (Theodore & Miller, 2010). The boundary between the /g/ and /k/ categories, as measured from the identification test, did not differ between the two training groups. However, the results showed a robust influence of training on internal category structure. The range of VOTs rated most prototypical of Joanne’s speech depended on how listeners had heard Joanne say /k/ during training. The VOTs rated “best” for Joanne occurred at shorter values for listeners who heard Joanne produce short-VOTs during training compared to listeners who heard Joanne produce long-VOTs during training. These results demonstrate that listeners begin to accommodate talker-specific phonetic variation at the earliest stages of mapping between the acoustic signal and linguistic representation. Moreover, this finding that listeners adjusted internal category structure but not the phonetic boundary provides additional evidence that these two aspects of phonetic categories can be independently affected by contextual variation and, with respect to earlier work (e.g., Eisner & McQueen, 2005), suggests that the ways in which a listener accommodates for talker-specific phonetic variation is very much dependent on the nature of a talker’s characteristic productions.

Introduction

The acoustic signal of speech simultaneously provides information about who is speaking and what is being said. That is, the same acoustic signal allows listeners to identify voices and linguistic content. Many findings, described in detail below, have indicated that listeners integrate these two types of information during speech processing. Specifically, researchers have shown that experience with a talker's voice facilitates linguistic processing of the acoustic signal, resulting in faster word recognition (Clarke & Garrett, 2004) and increased intelligibility of speech (Nygaard et al., 1994). Much of the research on talker familiarity has focused on higher levels of processing, such as word recognition. However, many recent studies suggest that listeners begin making talker-specific adjustments at the earliest stage linguistic processing (i.e., consonant and vowel identification). Below, we first review the findings demonstrating benefits of talker familiarity on language comprehension, highlighting findings from studies examining encoding of talker-specific phonetic variation in memory. We then present relevant background information on speech sound representations and review the findings that have shown that such representations are sensitive to variability in speech production, including variability that stems from idiosyncratic differences in production across individual talkers. The Introduction concludes by outlining the design and predictions of the current work.

Describing how listeners recover the segmental structure of language in spite of acoustic variability for individual consonants and vowels has been a primary goal of speech perception research. Many factors such as gender (Byrd, 1992), dialect

(Byrd, 1992), and vocal tract length (Peterson & Barney, 1952) contribute to the acoustic variability that exists in productions of the same segment. Even for a given talker, factors such as speaking rate (Miller & Liberman, 1979) and phonetic context (Liberman et al., 1961) create variability in the acoustic signal of a given consonant or vowel. Moreover, acoustic variability can be attributed to idiosyncratic variations in pronunciation that are characteristic of individual talkers. Such talker-specific phonetic variability has been shown for different classes of speech sounds, including fricatives (Newman et al., 2001), stops (Allen et al., 2003; Theodore et al., 2009), and vowels (Peterson & Barney, 1952). All of these sources of variability create the situation where there is no one-to-one mapping between the acoustic signal and a given word, or even a given consonant or vowel. However, despite this invariance, listeners are somehow able to accurately and seamlessly map the acoustic signal to phonetic segments without disruption in their understanding of the linguistic information. The goal of the current work is to contribute to a theoretical explanation of this process, focusing on variability associated with individual talkers' characteristic productions of stops.

Findings in the memory literature have shown that talker-specific variation in the acoustic signal of speech is stored in memory (e.g., Goldinger, 1996; Goldinger, 1998). In memory tasks, listeners show heightened recognition memory for words when talker is held constant on successive presentations of words compared to when talker varies. Other work in this domain has highlighted specific aspects of talkers' voices that are stored in memory, including emotional state, fundamental frequency and intonation patterns (Church & Schacter, 1994; Nygaard, Burt, &

Queen, 2000). Collectively, these findings demonstrate that one way listeners accommodate talker-specific phonetic variability is to retain it in memory, which raises the possibility that this information could be used to customize speech processing for individual talkers.

Indeed, researchers have demonstrated that listeners receive enhanced recognition of linguistic content when they are familiar with a particular talker's voice. Nygaard et al. (1994) investigated word intelligibility in noise of familiar and unfamiliar talkers. During training, listeners were trained to recognize a set of talkers over a period of nine days. Following training, listeners were tested with a list of words in the presence of noise and were asked to transcribe what they heard. Listeners who heard the list of words produced by the talkers presented during training demonstrated increased transcription accuracy compared to listeners who heard the list produced novel talkers. These findings indicate that listeners encoded and retained talker-specific speech patterns, and used this information to facilitate language comprehension.

In a later study, Bradlow and Pisoni (1999) replicated the above finding. However, in contrast to the talker exposure time of nine days as in Nygaard et al. (1994), listeners in this study demonstrated the effects of talker familiarity within the course of one testing session. The authors presented listeners with four blocks of word lists; some listeners heard the same talker over the four blocks and other listeners heard different talkers over the four blocks. To examine the effect of talker exposure, the authors measured intelligibility over time by comparing listeners' word transcription accuracy in the first and fourth block of test items. They found that overall transcription scores in the fourth block were significantly higher compared to the first block when talker voice was held constant, indicating that intelligibility increased as talker exposure also increased. More importantly, this study demonstrated that talker familiarity effects were evident after a short exposure time, indicating that the process of encoding talker-specific speech characteristics occurs fairly quickly. In fact, talker familiarity may even occur with as little exposure as two to four sentence-length utterances (Clarke & Garrett, 2004).

Taken together, findings from the memory and word recognition literature suggest that listeners accommodate talker-specific variability by using it to customize spoken language processing on a talker-by-talker basis. This effect has largely been demonstrated in the literature through word recognition studies. However, prior to word recognition, listeners must first process individual sounds by mapping them onto phonetic categories (e.g. McClelland et al., 1986). This raises the question that talker-specific processing might occur even earlier in the processing stream.

Most models of spoken language processing posit that listeners first map the acoustic signal onto phonetic categories prior to accessing lexical representations. Phonetic categories refer to representations for individual consonants and vowels that recognize variation in an acoustic-phonetic dimension as a single phonetic category. This organizational process begins in early infancy (e.g., Aslin, Pisoni, Jusczyk, 1983; Grieser & Kuhl, 1989; Eimas & Miller 1980) and is widely believed to account, in large part, for how listeners achieve perceptual constancy.

Some of the earliest empirical demonstrations of categorical processing concern how listeners process variation in voice-onset-time (VOT). VOT is an articulatory property of stop consonants and is defined as the time between the release of occlusion for the stop consonant and onset of vocal fold vibration for a subsequent vowel (Lisker & Abramson, 1964). This property can be measured acoustically and is an important distinction between voiced and voiceless stops. Figure 1 shows representative waveforms for /ba/ and /pa/, VOT for the voiced stop /ba/ is shorter than VOT for the voiceless stop /pa/. In English, voiced stops are produced with short VOTs and voiceless stops are produced with relatively longer VOTs (Lisker & Abramson, 1964). Evidence for categorical perception of VOT comes from studies that presented listeners with a continuum from /bi/ to /pi/ (Miller & Volaitis, 1989). The endpoints of the continuum presented VOTs that were typical of voiced and voiceless stops, but the intermediate members of the continuum consisted of fine-grain variations in VOT spanning the endpoint values. Listeners heard each member of the continuum and were asked to identify each token as beginning with /b/ or /p/. The results showed that /p/ responses were not linearly related to VOT duration; rather, listeners identified a range of VOTs as /b/, a different range of VOTs as /p/, and there was an abrupt discontinuity between the ranges. In other words, listeners appeared to have a VOT value that marked the boundary between the /b/ and /p/ categories. If a VOT was shorter than the boundary value, listeners perceived /b/. If a VOT was longer than the boundary value, listeners perceived /p/.

Findings such as these illustrate one property of phonetic categories; they have boundaries that mark how variation along a particular acoustic-phonetic dimension (e.g., VOT) is perceived. Other research has shown that phonetic categories, like other cognitive categories, also have a graded internal structure, in that not all members of a phonetic category are considered equally good members. Findings that have demonstrated this structure include presenting a /bi/ to /pi/ continuum to listeners and asking them to rate how “good” or prototypical each token represents the /p/ category. The results show that short-VOT tokens received the lowest ratings. This makes sense because they are the VOTs that are generally identified as /b/. Goodness ratings increased as VOTs become long enough to signal the /p/ category. However, as VOT continued to increase past the typical range of VOTs observed in speech production, goodness ratings systematically decreased. This occurs because VOTs begin to sound like extreme or highly aspirated versions of /p/, which are still unambiguously categorized as /p/, but not representative of how /p/ is typically produced. In other words, goodness rating tasks show that not all VOTs are considered equally good members of the /p/ category (Miller & Volaitis, 1989). Taken together, there is evidence that phonetic categories can be described by two important characteristics, boundaries between categories and internal category structure.

Research has shown that phonetic categories demonstrate functional plasticity such that the precise boundary and best exemplar region of a category shift as a consequence of systematic variation in the speech signal. Consider the example of speaking rate. In speech production, VOTs systematically increase as speaking rate slows (Miller & Volaitis, 1989). In speech perception, both the voicing boundary and the best exemplar region of /p/ are located at longer VOTs for a slow compared to a fast speaking rate (Miller & Liberman, 1979; Miller & Volaitis, 1989). In other words, listeners accommodate variability in the speech signal by shifting perceptual categories to reflect systematic patterns in speech production. It has been shown that – for phonetic category boundaries – these adjustments may be talker-specific (Eisner & McQueen, 2005; Kraljic & Samuel, 2005, but see Kraljic & Samuel, 2007). Eisner and McQueen (2005) tested two groups of listeners. Both groups were given exposure to an ambiguous sound midway between /f/ and /s/. In a lexical decision training phase, one group heard the ambiguous sound in the context of /f/-biased words such that if the sound was perceived as /f/, then the word would be considered a real word, whereas if perceived as /s/, then it would not be a real word (e.g., *effective*). The other group heard the same ambiguous sound in the context of /s/-biased words, where interpreting the ambiguous sound as /s/ would yield a real word but interpreting the sound as /f/ would not (e.g., *essential*). After this training, listeners were tested in a phonetic categorization task where they were presented with tokens along a continuum of [ɛf]–[ɛs] and asked to label the sound as "F" or "S;" the continuum was presented in the same voice as heard during training or in a different voice than heard during training. The results showed that when the test voice matched the training voice, listeners adjusted the [ɛf]–[ɛs] in line with their experience during training. However, no such boundary adjustment was observed when tested on the novel talker's voice. These results indicate that listeners used lexical information to resolve the ambiguous segment and then shifted the phonetic boundary in order to optimize processing of that variation. Critically, the boundary adjustment was talker-specific. This finding indicates that listeners can dynamically customize segmental organization on a talker-specific basis, which may result in comprehension benefits at other levels of linguistic processing.

As described above, phonetic categories also have a graded internal structure. Researchers have also shown that listeners shift the internal structure of phonetic categories in response to contextual influences, but it has not yet been determined whether this shift occurs on a talker-specific basis. Listeners are sensitive to talker differences in phonetic properties of speech, including VOT, which raises the possibility that they may use this sensitivity to reorganize phonetic category space in line with a talker's characteristic productions (Allen & Miller, 2004; Theodore & Miller, 2010). Examining sensitivity to talker differences within a given phonetic category provides a fundamental complement to findings, like that described above, that examine how listeners accommodate talker-specific productions that are ambiguous and thus fall near a phonetic category boundary. It may be the case that listeners incorporate a talker's characteristic productions in the same way for ambiguous versus clearly defined category members. However, it could also be the case that the type of adjustment listeners make depends on the nature of a talker's characteristic productions. In other words, one possibility is that listeners will show talker-specific boundary adjustments when adjusting for ambiguous productions and well-defined characteristic productions. An alternative is that listeners will show talker-specific boundary adjustments when adjusting for ambiguous productions and will show talker-specific internal category structure when adjusting for well-defined productions. Addressing these alternatives will provide critical information towards a theoretical account of speech perception that describes how listeners integrate talker and linguistic variability in the course of language comprehension.

The current work tests the hypothesis that listeners accommodate talker-specific phonetic detail by shifting phonetic category boundaries and internal category structure in line with a talker's characteristic productions. In training phases, we expose listeners to the speech of two talkers, fictitiously named "Joanne" and "Sheila." During training, listeners hear both voiced-initial and voiceless-initial tokens (i.e., *gain* and *cane*). However, here we manipulate characteristic VOTs such that for one group of listeners, Joanne produces /k/ with short VOTs compared to Sheila who produces /k/ with relatively longer VOTs. The other group of listeners hears the opposite pattern; Joanne produces /k/ with long VOTs relative to Sheila who produces /k/ with short VOTs. In all cases, both the short- and long-VOT variants of *cane* are unambiguously perceived as members of the /k/ category. Listeners are then tested in three ways: one test examines if characteristic VOT production is retained in memory, a different test examines if listeners shift the VOT voicing boundary as a consequence of exposure during training, and a third test examines if listeners shift the internal structure of /k/ as a consequence of training.

Based on previous research, we predict that listeners will encode talker-specific VOT in memory (Theodore & Miller, 2010). If listeners adjust phonetic boundaries to accommodate a talker's characteristic productions that are clearly defined category members as they do for productions that are ambiguous between categories, then we will observe a difference in the VOT voicing boundary between the two training groups. If listeners adjust internal category structure to be centered on a talker's characteristic productions, then we predict that the range of VOTs rated most prototypical will differ across the two training groups. These results will inform the perceptual mechanisms underlying listeners' ability to accommodate talker-specific phonetic detail.

Methods

Participants

Thirty-four participants were recruited from the University of Connecticut community. Participants were native monolingual speakers of English between 20-22 years of age with no history of speech, language or hearing impairment. Participants passed a 20 dB HL screen for hearing ability at 500 Hz, 1000 Hz and 2000 on the day of testing. The participants were randomly assigned to one of two training groups; half was assigned to the J-SHORT/S-LONG training group and the other half was assigned to the J-LONG/S-SHORT training group. The participants received monetary compensation for their participation in the study. As described below, high performance on talker identification and phonemic identification were required for inclusion in the data set. Because of these criteria, six participants were excluded from data analysis.

Stimulus creation

The stimuli consisted of two VOT continua, a continuum from gain to cane produced by two talkers with perceptually distinct voices. Creation of the continua follows procedures outlined in Theodore and Miller (2010). To sum, the continua were based on natural productions of the voiced-initial endpoint *gain*. Two female monolingual speakers of English were recorded producing many repetitions of these words (along with many fillers) and one repetition of each was selected such that word duration was approximately equivalent and the repetitions were of high acoustic quality (e.g., free from artifact). The selected *gain* tokens were equated for duration (568 ms) and a cosine ramp was applied to the final 30 ms of each token in order to simulate the naturally-occurring decrease in amplitude at word-offset.

A synthesized version of the selected *gain* tokens was created using LPC-based speech synthesis software (Analysis Synthesis Laboratory, Kay PENTAX) and this token served as the voiced-initial endpoint of each continuum, respectively. To create successive steps on each continuum, parameters of the LPC analysis were modified on a frame-by-frame basis (each frame corresponds to one vocal fold cycle) to replace the periodic source with a noise source and to scale peak amplitude by a factor of .22. After adjusting these parameters, a new token was synthesized based on the new parameters and the cycle was repeated. This procedure yielded, for each continuum, a series of tokens that incrementally increased in VOT in approximately 4 ms steps while maintaining constant word duration and filter characteristics of the original token. As described below, subsets of these continua were used as training and test stimuli.

Training stimuli

From each continuum, five tokens were selected for use during training: the voiced-initial endpoint, two tokens from the short-VOT voiceless region, and two tokens from the long-VOT voiceless region. VOTs of the selected tokens are shown in Table 1. The two short-VOT and two long-VOT voiceless tokens were selected such that they were two steps apart on the continuum. The short -and long-VOT variants were chosen such that they had the maximum difference in VOT yet the short-VOT variant was not so short that it fell in the ambiguous voiced/voiceless area, and the long-VOT was not so long that it was considered too extreme of a voiceless exemplar. The VOTs of the short-VOT and long-VOT tokens were equivalent across the two talkers. In order to equate the number of voiced and voiceless trials presented during training, a copy of the selected voiced-initial tokens was created. In order to eliminate a potential amplitude-based confound (see Theodore & Miller, 2010), two amplitude versions of the selected tokens were created, one corresponding to the root-mean-square (RMS) amplitude of the short-VOT voiceless tokens and one corresponding to the RMS amplitude of the long-VOT voiceless tokens. In total, 32 tokens were selected for use as training stimuli (2 voiced X 2 voiceless X 2 talkers X 2 amplitudes).

These stimuli were arranged into two different sets for using during the training phases. The J-SHORT/S-LONG set consisted of the voiced-initial tokens from both talkers, Joanne's short-VOT voiceless tokens, and Sheila's long-VOT voiceless tokens. The J-LONG/S-SHORT set consisted of the voiced-initial tokens from both talkers, Joanne's long-VOT voiceless tokens, and Sheila's short-VOT voiceless tokens.

Test Stimuli

Two-alternative forced choice (2AFC) test. All test stimuli were drawn from Joanne's continuum. Stimuli for the 2AFC consisted of pairs of stimuli, a short-VOT variant and a long-VOT variant, separated by 750 ms of silence. Recall that for the training stimuli, short- and long-VOT variants were selected such that they were two steps apart on the continuum. Each stimulus pair for the 2AFC test was formed using the intermediate tokens. Four pairs were created using the two amplitude variants of the selected short-VOT and long-VOT test token; amplitude was held constant on a given pair and half presented the short-VOT token first, with the other half presenting the long-VOT tokens first. VOTs of the selected test tokens are shown in Table 2.

Identification and goodness tests. The same stimuli were used for the identification and goodness rating tests and are described in Table 2. These stimuli were drawn from Joanne's continuum and consisted of 24 tokens spanning the VOTs of 25 ms to 183 ms. These VOTs represent the range of VOTs presented during training and thus span VOTs of the voiced-initial tokens and the long-VOT voiceless tokens; however, none of the voiceless-initial tokens used at test were physically identical to those presented during training. RMS amplitude was held constant across the selected tokens. Step size of the first 12 tokens was 4-5 ms and step size of the last 12 tokens was 8-10 ms.

Procedure

As stated above, the participants were randomly assigned to either the J-SHORT/S-LONG training group or the J-LONG/S-SHORT training group. The only difference across the training groups concerned the stimuli presented during training. The overall procedure required listeners to participate in training phases and test phases, described in detail below. All testing took place in a sound-attenuated booth. Listeners were seated at a table that held a computer monitor and a button box. Visual stimuli were presented on the monitor and auditory stimuli were presented via headphones. All responses were collected via button box. Participants were given the option to take breaks throughout the experiment. Participants were also instructed to always respond to every trial and encouraged to make their best guess if they were unsure of how to respond. The entire protocol took approximately 2 hours to complete. Below we describe the procedure for the training and test phases, and then we describe the overall procedure.

Training phases

Stimuli presented during training were the lists designed for each training group. On each trial, the participant heard an auditory stimulus consisting of Joanne or Sheila saying either *gain* or *cane*. Participants were asked to indicate whether they heard Joanne's voice or Sheila's voice and if they heard *gain* or *cane*. The participants indicated their responses by pushing one of four buttons labeled "Joanne G," "Joanne K," "Sheila G," and "Sheila K." Feedback was provided for the talker choice only on the computer monitor. A 750 ms pause occurred between the offset of the auditory stimulus and visual feedback, which showed "YES" for correct responses and "No. That was Joanne." or "No. That was Sheila." for incorrect responses. Visual feedback remained on the screen for 1000 ms. Each trial was separated by a 1500 ms pause measured from the offset of the visual feedback. Each training block consisted of three randomizations of the training stimuli described above, for a total of 48 trials.

2AFC test

Each 2AFC test phase consisted of two randomizations of the 4 test pairs created for Joanne's voice. The stimulus pairs for this test phase thus consisted of the short-VOT and long-VOT variants of Joanne's *cane.* Participants were instructed to choose which item in each pair sounded more characteristic of Joanne based on their previous experience with her voice. Participants indicated their choice by pushing a button labeled "1" for the first member of the pair or "2" for the second member of the pair. Each trial was separated by a 1500 ms pause measured from the listener's response. No feedback was provided at test.

Identification test

The purpose of this test was to identify the point along the VOT continuum that marked where listeners marked the voicing boundary. Stimuli thus consisted of the selected members of Joanne's VOT continuum. Each identification test phases consisted of one randomization of the 24 test tokens. Participants were instructed to listen to each stimulus and indicate whether they heard *gain* or *cane*. Participants indicated their choice by pushing a button labeled "G" for *gain* or a button labeled "K" for *cane*. A 1500 ms pause separated each trial measured from the listener's response. No feedback was provided at test.

Goodness rating test

The same tokens used in the identification test were also used in this goodness test. One randomization of the 24 test stimuli was presented in each test phases. For each stimulus, participants were instructed to rate each token for goodness as /k/ based on their previous experience with Joanne's voice. Listeners used a 1-7 scale to respond, with 7 indicating the best exemplar. Instructions indicated that tokens that sounded similar to *gain* should receive very low ratings, while tokens that seemed to match exactly how Joanne said *cane* should receive the highest ratings. Seven buttons on a button box were labeled from 1 through 7 accordingly and participants were instructed to indicate their rating by pressing appropriate button. A 1500 ms pause separated each trial measured from the listener's response. No feedback was provided at test.

Experiment proper

The experiment began with a familiarization phase in which listeners were given the opportunity to learn the names of the talkers' voices. One randomization of the training stimuli was presented and the name of the talker for each stimulus appeared on the computer monitor. Listeners were instructed to listen and learn the names of the talkers; no responses were collected.

Following familiarization, listeners completed three blocks of training and test phases. Each block consisted of six alternations of training phases and test phases, blocked by the particular type of test. All listeners completed the 2AFC test first. Thus, listeners began by alternating between a training phase, then the 2AFC test phase, and then another training phase, and so on to the completion of six training and test phases. Following the 2AFC test, listeners completed the other two tests similarly, with order of the identification and goodness rating tests counterbalanced within each training group.

Results

Training

Performance during training was analyzed separately for each training group and for each talker. Two measures of accuracy were calculated, one for talker identification and one for phonetic identification. Examining accuracy for talker identification allows us to examine if listeners learned the talker's voices. Examining accuracy for phonetic identification allows us to examine if the stimuli presented during training were perceived as intended (i.e., that the short-VOT voiceless tokens were perceived as /k/ and not as /g/). For each listener, mean percent correct talker identification was calculated by collapsing across the trials presented during the six training phases for a particular test. A response was considered correct if the talker was identified correctly, even if the phonetic decision was incorrect. If a subject failed to meet the criterion of 80% correct of higher, he or she was excluded from further analysis. One participant was excluded for this reason. Mean percent correct phonetic identification was similarly calculated for each listener, and five participants were excluded because they failed to meet the accuracy criterion.

Mean performance across listeners for all training phases is shown in Figure 2. Performance was near ceiling for both training groups and for both talkers, for both the talker and phonetic decisions (mean > 95% in all cases). These results indicate that the listeners learned the talkers' voices and perceived the VOT variants as intended.

Test

Two-alternative forced choice

Performance during the 2AFC test sessions was analyzed separately for the J-SHORT/S-LONG and J-LONG/S-SHORT training groups. Recall that on each trial during test, listeners chose either a short-VOT variant of *cane* or a long-VOT variant of *cane*. For each subject, percent long-VOT responses was calculated by collapsing across the eight pairs within each test block and then collapsing across the six test blocks. Percent long-VOT responses was used as the dependent measure to ease comparison to earlier work (Theodore & Miller, 2010); analyzing percent short-VOT responses would have been appropriate, however analyzing both percent short-VOT and long-VOT responses is redundant given that they must sum to 100.

Figure 2 shows mean percent long-VOT responses for the two training groups. As can be seen in this figure, percent long-VOT responses was higher for the J-LONG/S-SHORT training group compared to the J-SHORT/S-LONG training group, in line with previous exposure to Joanne's voice during training. That is, listeners who heard Joanne produce *cane* with long VOTs during training chose more long-VOT variants of *cane* at test compared to listeners who heard Joanne produced *cane* with short VOTs during training. The difference in percent long-VOT responses between the two groups was statistically reliable [t(26) = 3.44, p = .002], and indicates that performance during training guided performance at test, as was predicted by previous findings (Allen & Miller, 2004; Theodore & Miller, 2010).

Identification

Performance during the identification test sessions was analyzed separately for the J-SHORT/S-LONG training group compared to the J-LONG/S-SHORT training group. For each listeners, percent /k/ responses was calculated for each step of the VOT continuum presented at test by collapsing across the six test sessions. Mean performance across the listeners in shown in Figure 3. Consider first performance for the J-SHORT/S-LONG training group. Percent /k/ responses are near zero for the shortest VOTs of the continuum and near ceiling for the longest VOTs of the continuum, and the there is an abrupt discontinuity between the two ranges of VOTs. This pattern of performance indicates that listeners processed the VOT continuum categorically, as predicted (e.g., Volaitis & Miller, 1992), and the same pattern is observed for the J-LONG/S-SHORT training group. However, our question concerns the degree to which the boundary between /g/ and /k/ responses differences between the two training groups. Inspection of the figure suggests that the boundary of the J-SHORT/S-LONG training group is located at a slightly longer VOT compared to the J-LONG/S-SHORT training group.

To examine the statistical significance of this displacement, a voicing boundary was calculated for each listener as follows. Probit analyses were used to fit an ogive function to percent /k/ responses for a given subject. This process thus fit responses to a cumulative normal distribution. The mean of this distribution was calculated, defined as the VOT (ms) corresponding to 50% of the cumulative normal distribution. Thus, the mean of the curve marks the VOT boundary where half of the responses fall into the *gain* category and the other half of the responses fall into the *cane* category. This metric was used to quantify the voicing boundary for each listener. For all listeners, the fitted curve was an excellent fit to the identification data as indicated by *r*, which ranged from 0.98 to 1.00 across the participants. This process is illustrated in Figure 4, which shows a representative function from one of the listeners. Figure 5 shows the mean boundary across listeners for the two training groups. Though there is a numerical difference between the two groups, the difference in voicing boundary between the two training groups was not statistically reliable [t(13.840) = 1.81, p = .092].[1] These results indicate that experience with Joanne's voice during training did not influence performance during the identification test sessions. In other words, contrary to earlier work showing that listeners accommodate a talker's ambiguous production by shifting phonetic boundaries, here we found no evidence that listeners make such boundary adjustments to accommodate a talker's unambiguous characteristic productions.

Goodness ratings

Performance during the goodness rating test sessions was analyzed separately for the J-SHORT/S-LONG training group compared to the J-LONG/S-SHORT training group. For each listener, mean goodness as /k/ was calculated for each step of the VOT continuum by collapsing across the six test sessions. Mean performance across the listeners in shown in Figure 6. For both training groups, mean goodness ratings were extremely low for the short VOT tokens. This is as

[1] Levine's test for equality of variances indicated that the two groups violated the homogeneity of variance assumption of the independent t-test [F = 12.49, p = .002] and thus the degrees of freedom were adjusted for this comparison following the Welch-Satterthwaite method.

expected given that these are VOTs that in the identification test were perceived as /g/; thus, tokens perceived as /g/ would be rated as very poor exemplars of /k/. For both training groups, mean goodness ratings increase as does VOT, reflecting the fact that as VOT increases these tokens are now actually perceived as /k/. However, as can be seen in the figure, the two training groups do not show identical goodness functions, particularly for the range of VOTs presented during training. The goodness ratings peak at shorter VOTs for the J-SHORT/S-LONG training group compared to the J-LONG/S-SHORT training group. In fact, mean goodness ratings begin to decrease for the J-SHORT/S-LONG training group before goodness ratings in the J-LONG/S-SHORT training group have reached their peak. This pattern suggests that perceived goodness of Joanne's /k/ differed as a consequence of previous experience with her voice.

To quantify the statistical significance of this pattern, a best exemplar range was calculated for each listener using the conventions outlined in Allen and Miller (2003). First, the peak rating for a particular listener was identified. The best exemplar range was defined as the range of VOTs that fell within 90% of the peak rating. For example, if a listener had a peak rating of 7, the best exemplar region was considered as the range of VOTs that were given a rating of 6.3 and higher. The lower bound of the best exemplar region was calculated by determining the VOT value where ratings increased above 90% of the peak and the lower bound of the best exemplar region was calculated by determining the VOT value where ratings decreased below 90% of the peak. When the "90% of peak rating" criterion fell between obtained goodness ratings for consecutive tokens, linear interpolation was used to determine the VOT value that would have corresponded to the criterion. Figure 7 illustrates this process by providing a representative function and best exemplar region from one of the listeners. The horizontal lines above the goodness functions in Figure X show the mean best exemplar regions for each training group. Across listeners, the best exemplar region for the J-SHORT/S-LONG training group ranged from 88 ms to 144 ms and the best exemplar region for the J-SHORT/S-LONG training group ranged from 122 ms to 177 ms. Independent t-tests showed that the lower bound [t(12.948) = -3.65, p = .003] and the upper bound [t(22.816) = -4.95, p < .001] of the best exemplar regions were located at significantly shorter VOT values for the J-SHORT/S-LONG training compared to the J-LONG/S-SHORT training group.[2] This pattern indicates that experience with Joanne's voice during training guided performance at test; specifically, listeners adjusted internal category structure in line with Joanne's characteristic productions.

[2] Levine's test for equality of variances indicated that the two groups violated the homogeneity of variance assumption of the independent t-test for comparison of both the lower [F = 28.05, p < .001] and upper bounds [F = 4.56, p = .042] of the best exemplar region. Accordingly, the degrees of freedom were adjusted for these comparisons following the Welch-Satterthwaite method.

Discussion

The acoustic-phonetic signal of speech contains a lot of variability for individual speech segments. That is, there is no one-to-one mapping between the acoustic signal and the individual speech sounds. As discussed earlier, there are many sources that contribute to this variability including speaking rate (Miller & Liberman, 1979) and gender (Byrd, 1992). In addition, talkers have idiosyncratic patterns in speech production that give rise to talker-specific implementation of individual consonants and vowels (e.g., Newman et al., 2001). One such example is that talkers differ in their characteristic VOT production; some talkers have longer VOTs relative to other talkers (Allen et al., 2003; Theodore et al., 2009).

Despite this variability, listeners map the acoustic signal to a phonetic segment without disruption in comprehension of the linguistic message. Regarding talker-specific phonetic variability, there is growing evidence that listeners achieve such perceptual constancy by encoding talker-specific phonetic detail in memory. Indeed, talker familiarity has been shown to facilitate speech intelligibility (Nygaard et al., 1994; Bradlow & Pisoni, 1999) and decrease processing time (Clarke & Garrett, 2004).

Researchers have found that talker-specific encoding begins early in the processing stream at the phonetic level, prior to word recognition. Listeners are sensitive to talker differences in individual phonetic properties of speech that are used to identify individual consonants and vowels, including VOT (Theodore & Miller, 2010). Moreover, research has shown that listeners make adjustments to phonetic category boundaries in light of talker-specific differences in speech production (Eisner & McQueen, 2005; Kraljic & Samuel, 2005).

As reviewed in the Introduction, phonetic categories are marked not only by boundaries, but they also exhibit a graded internal structure. This internal structure has been shown to shift to accommodate systematic variation in speech production, including that associated with changes in speaking rate (Miller & Volaitis, 1989). However, previous research has not examined whether such reorganization of internal category space is applied on a talker-by-talker basis, and this was the primary focus of the current work.

Our results provide additional evidence that listeners begin talker-specific processing of the acoustic speech signal at the earliest stages of comprehension when they extract individual consonants and vowels from the speech stream. Listeners were differentially exposed to a talker's characteristic productions in training phases. Performance at test showed that listeners encoded these differences in memory and adjusted the internal category structure to reflect previous experience with the talker's voice. The results did not indicate, however, that the phonetic category boundary was influenced by exposure during training.

In contrast to earlier findings, the results from the current work do not provide evidence to support the notion that listeners adjust phonetic category boundaries to accommodate talker-specific phonetic variation. Here, we consider two possible explanations for this discrepancy. The first is methodological in nature. The present study contained 28 participants, which may be too small of a sample to have statistical power to detect group differences. Previous research in these paradigms has used more participants and additional data for the current work is currently in progress. In addition, the number of test trials in our study compared to similar paradigms is smaller, which could contribute to decreased power as well.

The second consideration for this discrepancy is more theoretical in nature. In earlier work that provided evidence for phonetic category boundary adjustments as a consequence of talker exposure (Eisner & McQueen, 2005; Kraljic & Samuel, 2005), listeners were presented with an ambiguous token that fell on a category boundary. However, in the present study, listeners were presented with tokens that were unambiguous exemplars within the phonetic category /k/. It may be possible that how listeners make talker-specific adjustments to phonetic categories is contingent on the particular idiosyncratic production. For example, if a talker's production is ambiguous, the system may adjust by moving a category boundary so as to support lexical recognition. However, if a talker's production is clear and unambiguous, the system may not need to change the boundary in order to accommodate the talker's production. Rather, the listener may be able to customize segmental processing solely by a reorganization of internal category space. There is empirical support for a disassociation between functional plasticity of category boundaries and internal category space.

Recall that speaking rate and lexical status have both been shown to influence functional plasticity of stop voicing categories (Allen & Miller, 2003; Volaitis & Miller, 1992). Of these two contextual influences, only speaking rate creates systematic variation in the speech signal. That is, as speaking rate slows in production, so too do VOTs for word-initial stop consonants. No such effect of lexical status is observed for speech production; there is no systematic difference in VOT values for words (e.g., *beef*) and nonwords (e.g., *beace*). Both of these contexts influence perception of stop consonant voicing; however, only speaking rate causes a shift in both the boundary and internal category structure. That is, the voicing boundary and the best exemplar region are shifted towards longer VOTs for a slow compared to a fast speaking rate. Influences of lexical status are only observed at the boundary. This decoupling has been explained as the consequence of a perceptual system that has tight links to the acoustic signal of speech. Because lexical status does not come with a concomitant change in speech production, listeners do not modify internal category structure. In this vein, the talker-specific adjustments we observed in the current work may be the reflection of the precise acoustic-phonetic information we provided during training. Listeners in the J-SHORT/S-LONG training group heard Joanne produce /k/ with shorter VOTs compared to the J-LONG/S-SHORT training group; however, the VOT Joanne produced for /g/ was identical in the two groups. Thus, unlike the contextual influence of speaking rate, which affects voiced and voiceless stop consonants, the talker-specific productions in our work were limited to the voiceless category. For this reason, listeners were able accommodate the talker difference solely within the internal category space. This explanation makes the broad hypothesis that listeners will accommodate a talker's characteristic productions only to the degree that the acoustic signal requires them to. This hypothesis would predict perceptual accommodations for a talker's production that is ambiguous would be limited to the boundary region. This hypothesis is currently being tested in related experiments.

In moving forward, future research is aimed at confirming that the talker-specific adjustments to internal category structure are not limited to the word presented during training. Previous research has shown that these effects are not limited to words presented during training; rather, listeners encode this information for a phonetic category broadly. Thus we predict that in fact the shifts in internal category structure will be observed if listeners were tested on a novel word.

To sum, the results from the current work add to the body of evidence indicating that listeners begin to accommodate talker-specific phonetic variation at the earliest stages of mapping between the acoustic signal and linguistic representation, and thus may underlie, at least in part, talker familiarity effects observed at higher levels of linguistic processing.

References

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

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

Aslin, R. N., Pisoni, D. B., & Jusczyk, P. W. (1983). Auditory development and speech perception in infancy. In M. M. Haith & J. J. Campos (Eds.), *Carmichael's manual of child psychology: Vol. 2. Infancy and the biology of development (4th ed., pp. 573-687).* New York: Wiley.

Bradlow, A. R., & Pisoni, D. B. (1999). Recognition of spoken words by native and non-native listeners: Talker-, listener-, and item-related factors. *The Journal of the Acoustical Society of America*, *106*, 2074.

Byrd, D. (1992). Preliminary results on speaker-dependent variation in the TIMIT database. *Journal of the Acoustical Society of America, 92*(1), 593-596.

Church, B. A., & 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(3),* 521-533.

Clarke, C. M., & Garrett, M. F. (2004). Rapid adaptation to foreign-accented English. *The Journal of the Acoustical Society of America, 116(6),* 3647.

Eimas, P. D., & Miller, J. L. (1980). Contextual effects in infant speech perception. *Science*, *209*(4461), 1140-1141.

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

Goldinger, S. D. (1996). Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology-Learning Memory and Cognition*, *22*(5), 1166-1182.

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological review*, *105*(2), 251.

Grieser, D., & Kuhl, P. K. (1989). Categorization of speech by infants: Support for speech-sound prototypes. *Developmental Psychology*, *25*(4), 577.

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

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

Liberman, A. M., Harris, K. S., Eimas, P., Lisker, L., & Bastian, J., An effect of learning on speech perception: the discrimination of durations of silence with and without phonemic significance. *Language and Speech, 4(4),* 175-195.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, *74*(6), 431-461.

Lisker, L., & Abramson, A. S. (1964). A cross-language study of voicing in initial stops: Acoustical measurements. *Word*, *20(3),* 384-422.

McClelland, J. L., & Ellman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology, 18(1),* 1-86

Miller, J. L., & Liberman, A. M. (1979). Some effects of later-occurring information on the perception of stop consonant and semivowel. *Perception & Psychophysic*s, *25(6),* 457-465.

Miller, J., & Volaitis, L. (1989). Effect of speaking rate on the perceptual structure of a phonetic category. *Perception & Psychophysics, 46(6),* 505-512.

Newman, R. S., Clouse, S. A., & Burnham, J. L. (2001). The perceptual consequences of within-talker variability in fricative production. *The Journal of the Acoustical Society of America, 109(3),* 1181-1196.

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

Nygaard, L. C., Sommers, M. S., & Pisoni, D. B. (1994). Speech perception as a talker-contingent process. *Psychological Science, 5(1),* 42-46.

Nygaard, L. C., Burt, S. A., & 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(5),* 1228-1244.

Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the vowels. *The Journal of the Acoustical Society of America, 24(2),* 175-184.

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

Theodore, R. M., Miller, J. L., & DeSteno, D. (2009). Individual talker differences in voice-onset-time: Contextual influences. *The Journal of the Acoustical Society of America, 125(6),* 3974-3982.

Theodore, R. M., & Miller, J. L. (2010). Characteristics of listener sensitivity to talker-specific phonetic detail. *The Journal of the Acoustical Society of America, 128(4),* 2090-2099.

Volaitis, L. E. & Miller, J. L. (1992). Phonetic prototypes: Influence of place of articulation and speaking rate on the internal structure of voicing categories. *The Journal of the Acoustical Society of America, 92(2 Pt 1),* 723-735.

List of figures

horizontal line). VOTs that met or exceeded this criterion constituted the best exemplar region, shown here by the solid horizontal line. For this participant, the lower bound of the best exemplar region was 81 ms and the upper bound of the best exemplar region was 127 ms.

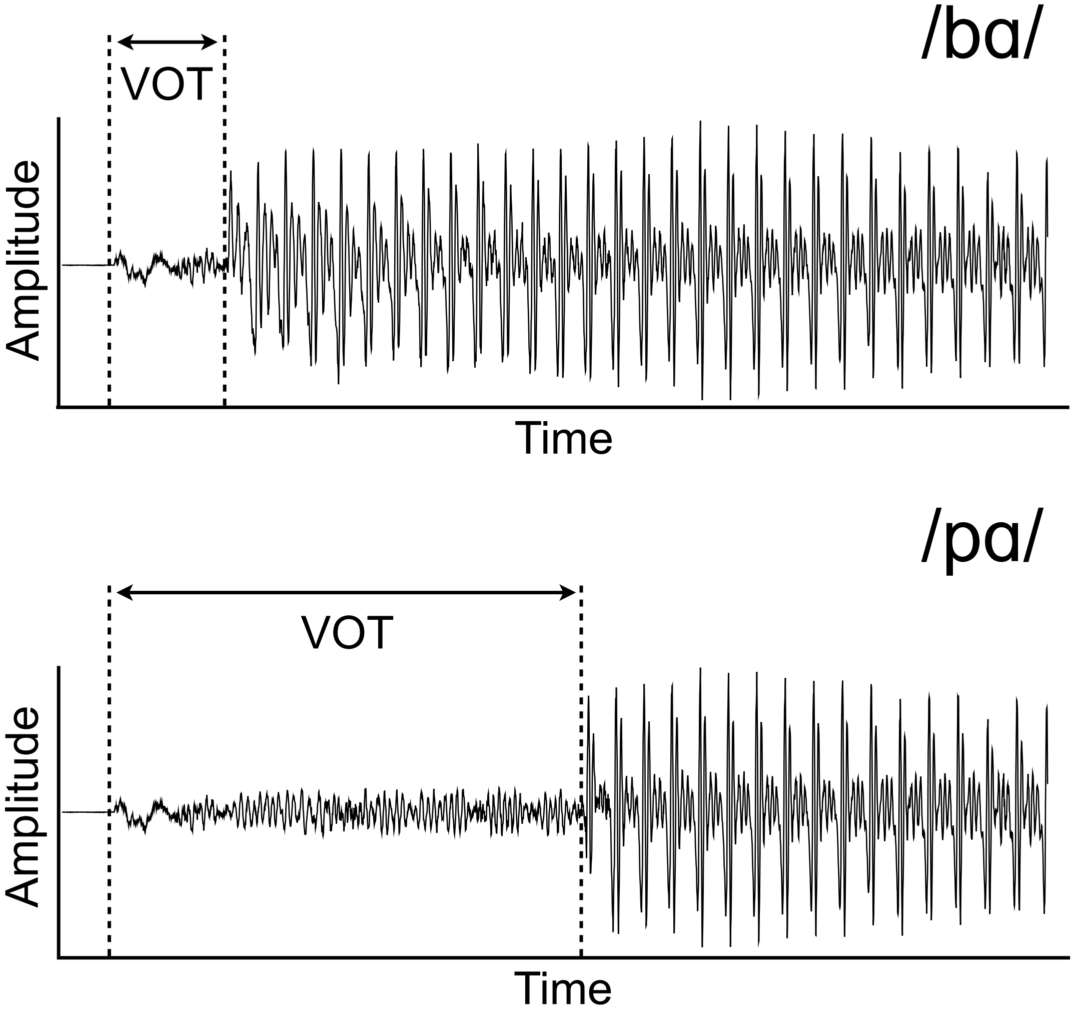

Figure 1. Representative waveforms showing voice-onset-time for a voiced stop (top panel) and a voiceless stop (bottom panel).

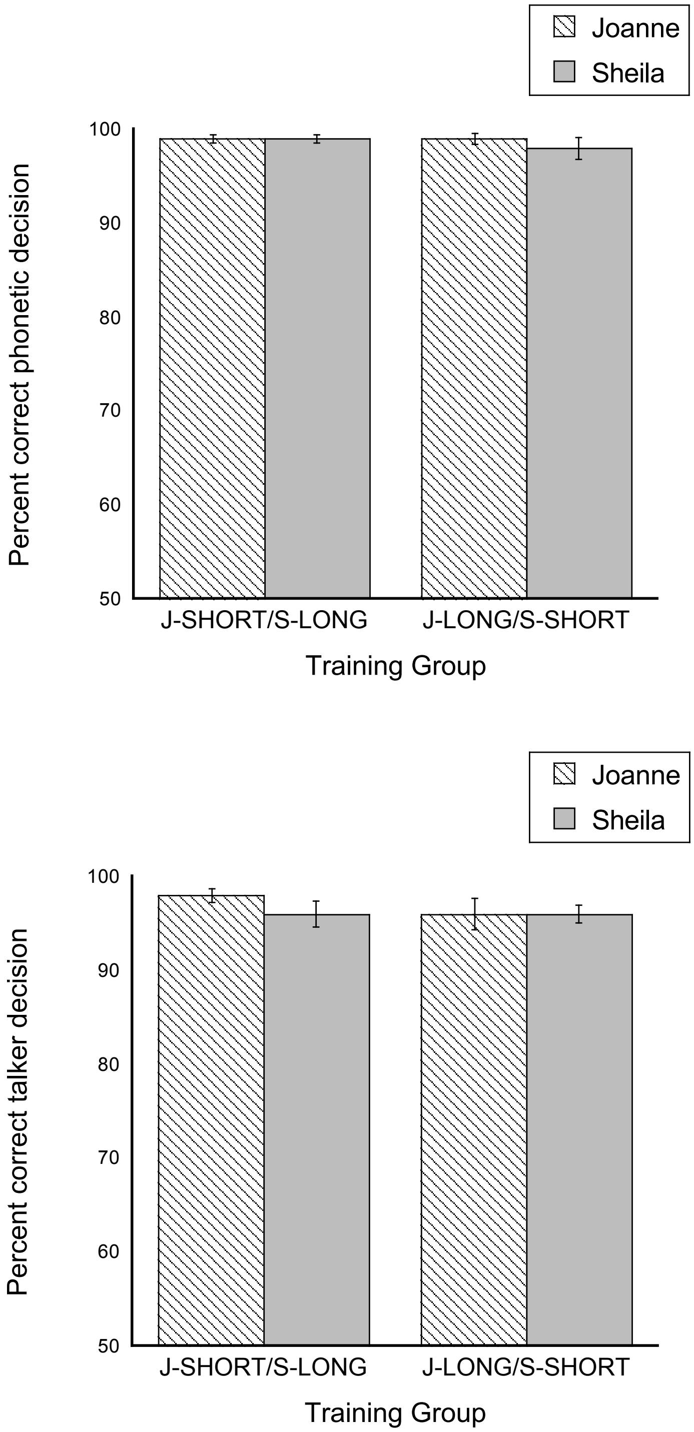

Figure 2. Mean percent correct phonetic identification (top panel) and talker identification (bottom panel) for the two talkers for the two training groups. Error bars indicate standard error of the mean.

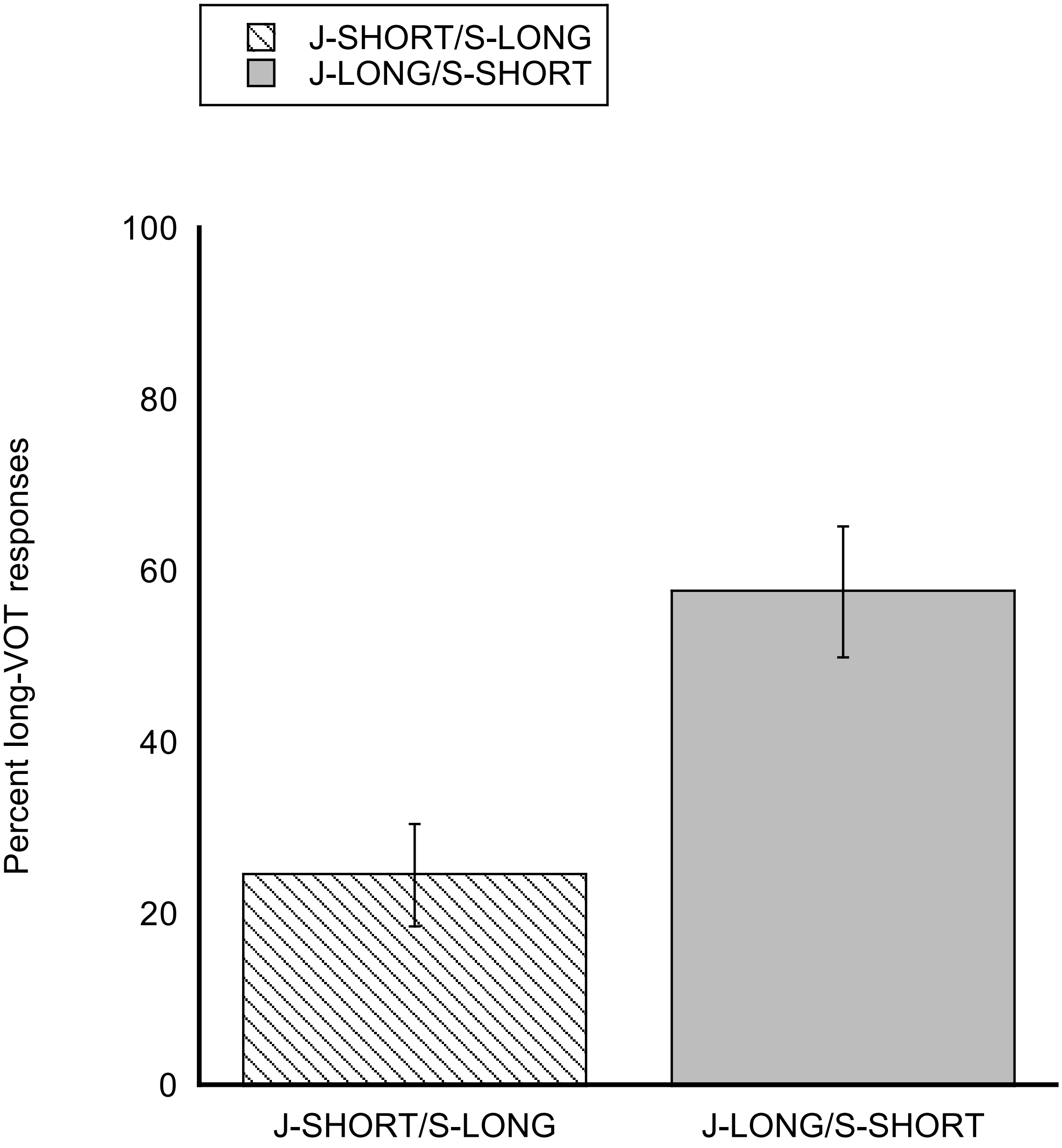

Figure 3. Mean percent long-VOT responses for the two training groups. Error bars indicate standard error of the mean.

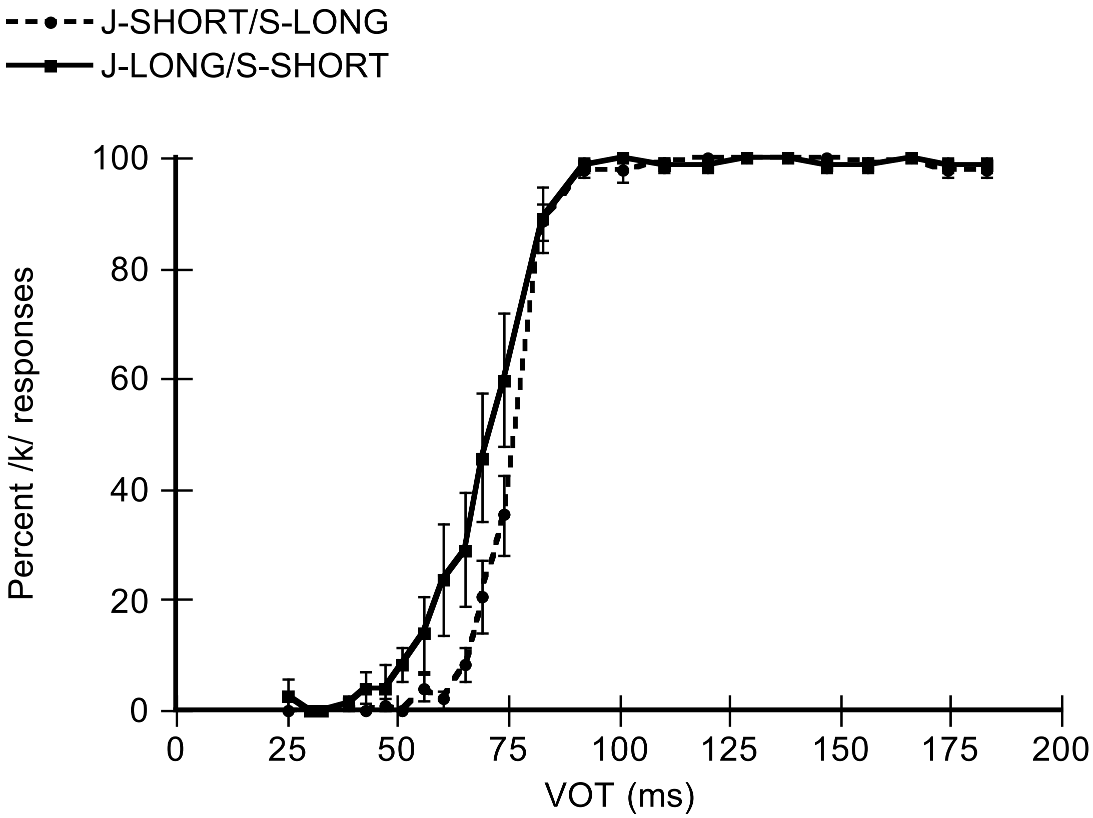

Figure 4. Identification functions for the two training groups. Mean /k/ responses are shown as a function of VOT (ms). Error bars indicate standard error of the mean.

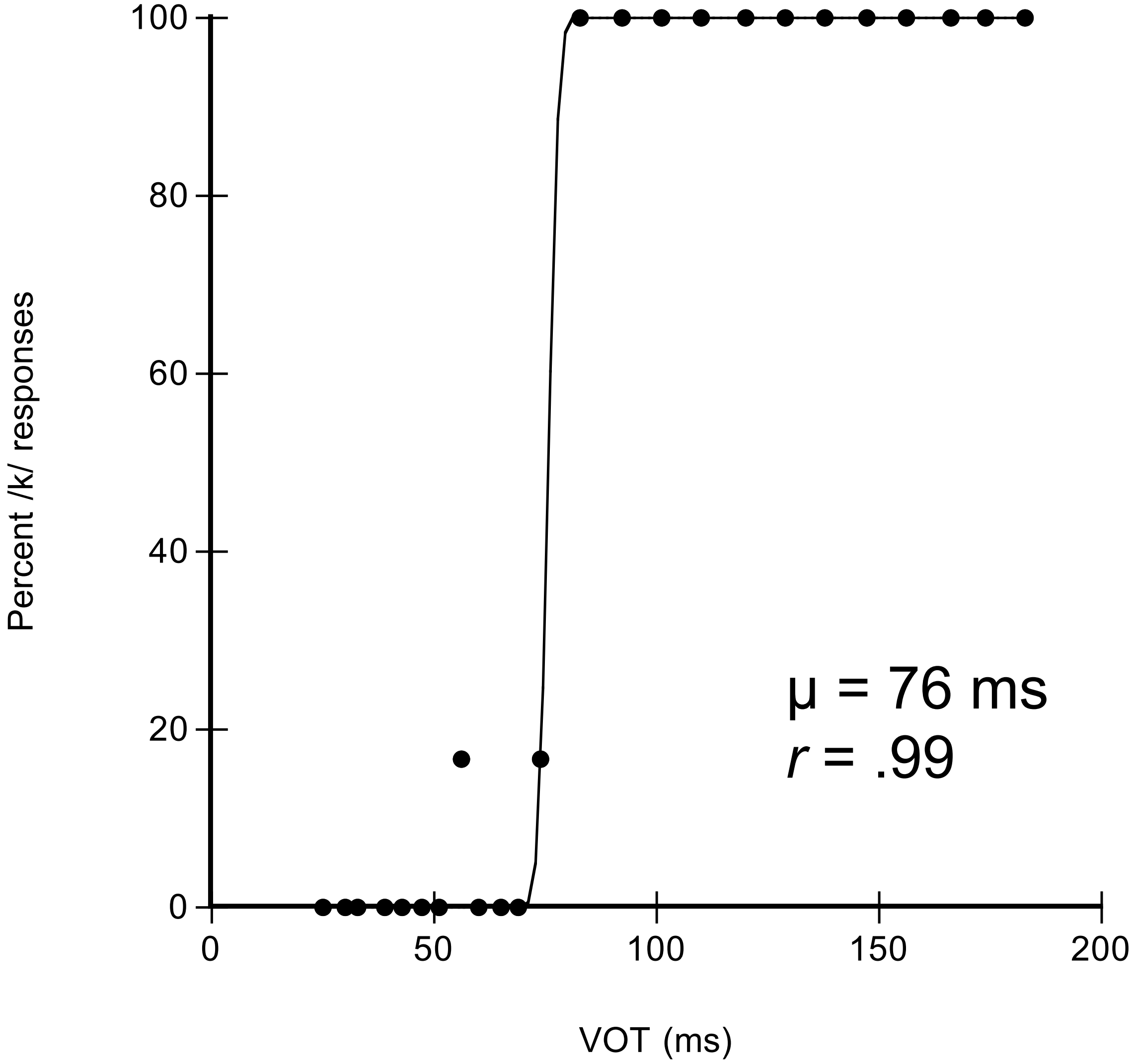

Figure 5. Representative identification function illustrating the use of probit analyses to determine the voicing boundary. Obtained data points are shown in filled circles and the line shows the fitted curve. The mean of the curve (μ) used as the boundary and goodness of fit (*r*) are shown.

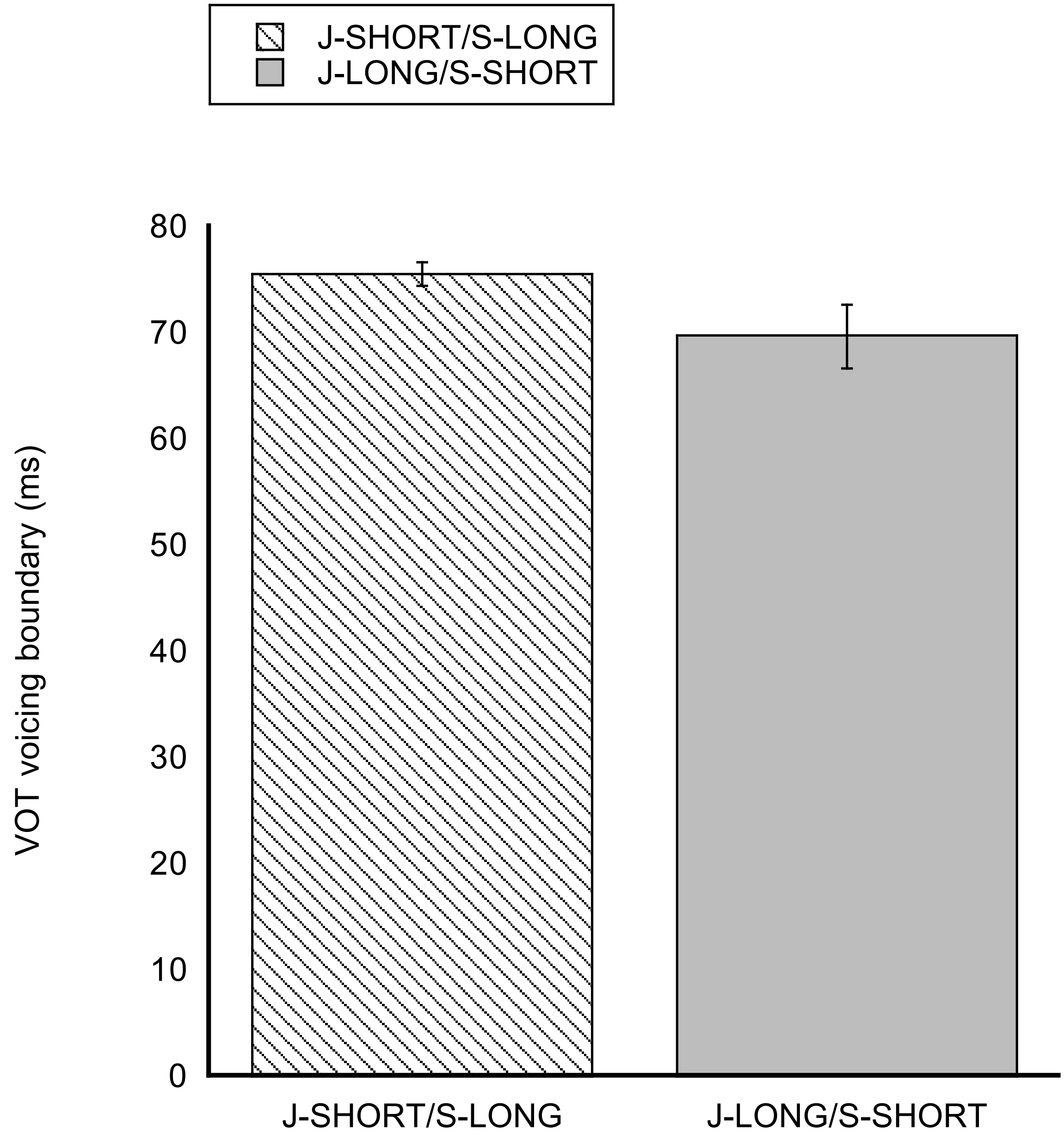

Figure 6. Mean VOT (ms) voicing boundary for the two training groups. Error bars indicate standard error of the mean.

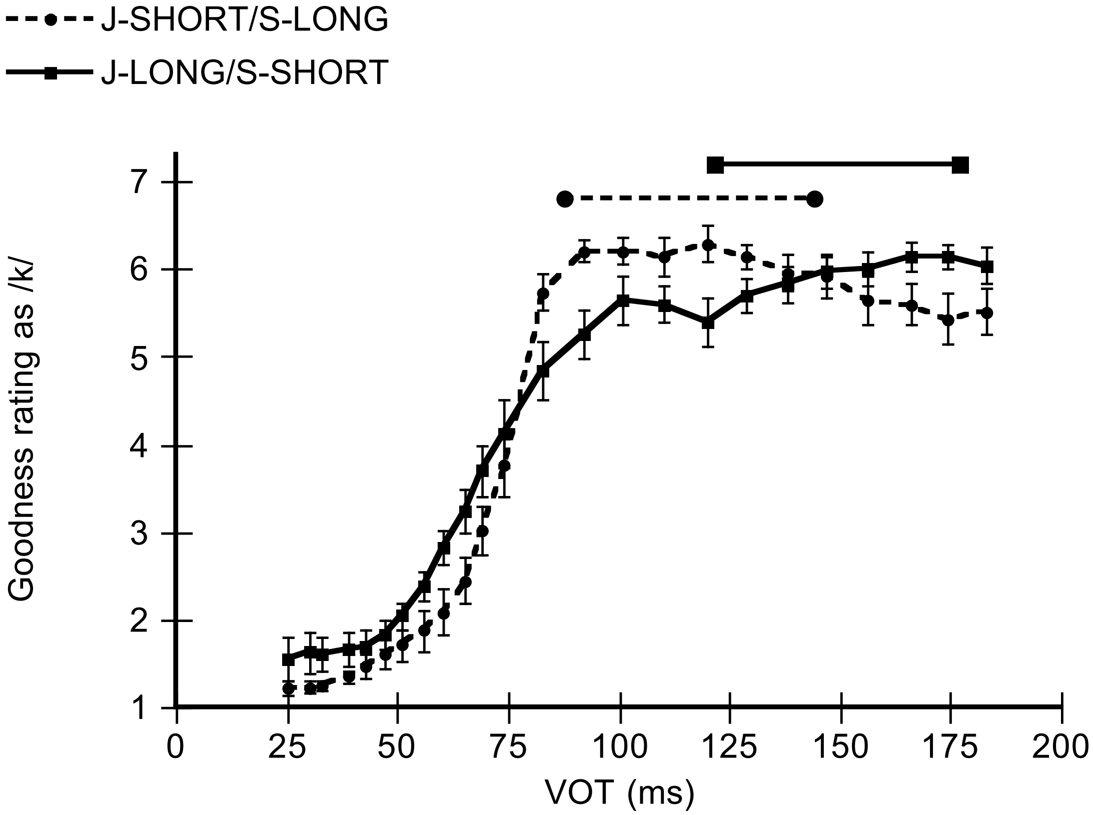

Figure 7. Mean good ratings for the two training groups. Error bars indicate standard error of the mean. The horizontal lines indicate the best exemplar regions.

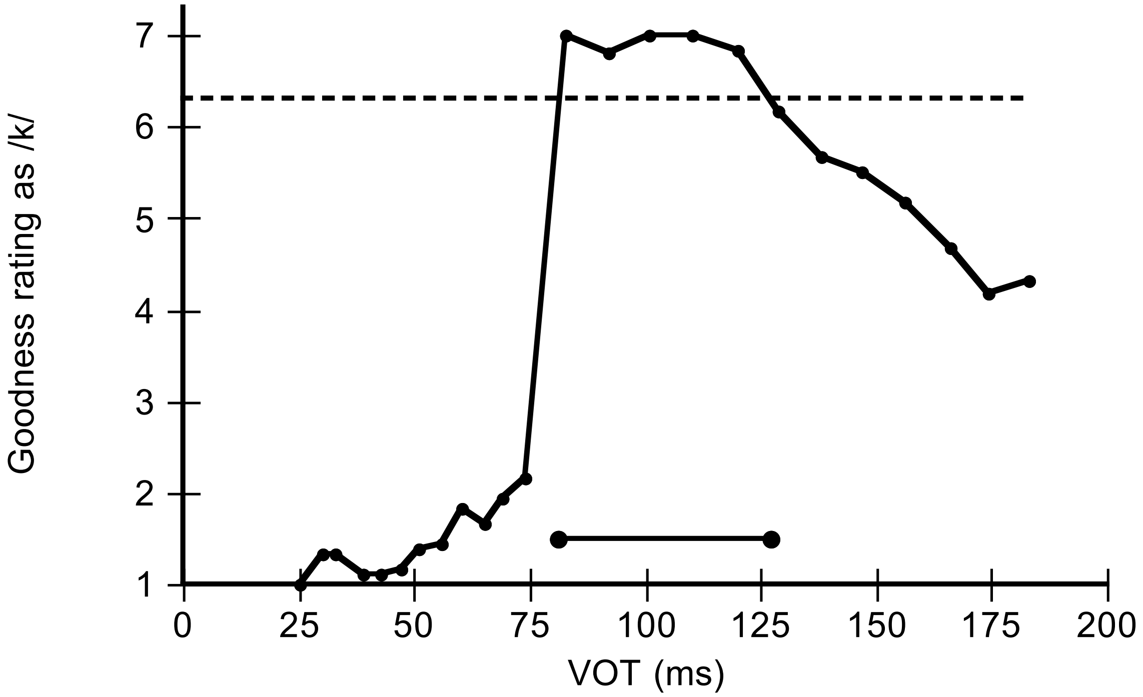

Figure 8. Representative goodness function to illustrate calculation of the best exemplar region. For this participant, the peak rating was 7 and the "90% of peak rating" criterion was 6.30 (shown by the dashed horizontal line). VOTs that met or exceeded this criterion constituted the best exemplar region, shown here by the solid horizontal line. For this participant, the lower bound of the best exemplar region was 81 ms and the upper bound of the best exemplar region was 127 ms.

List of tables

Table 1

VOT in milliseconds for the stimuli presented during training for the two training groups.

	Joanne		Sheila	
Training Group	Type	VOT (ms)	Type	VOT (ms)
J-SHORT/S-LONG	Voiced	22	Voiced	20
	Voiceless	69 78	Voiceless	172 181
J-LONG/S-SHORT	Voiced	22	Voiced	20
	Voiceless	170 179	Voiceless	172 181

Table 2

VOT in milliseconds for the stimuli presented during the test phases. Test stimuli were produced by Joanne and were the same for both training groups. Tokens in bold denote those used for the pairs in the 2AFC test.

Token	VOT (ms)
1	25
2	30
3	33
4	39
5	43
6	47
7	51
8	56
9	60
10	65
11	69
12	**74**
13	83
14	92
15	101
16	110
17	120
18	129
19	138
20	147
21	156
22	166
23	**174**
24	183