Talker-specific perceptual processing: Influences on internal category structure

Rachel M. Theodore

Department of Speech, Language, and Hearing Sciences, University of Connecticut, Storrs, Connecticut, United States

rachel.theodore@uconn.edu

Abstract

Research has shown that listeners accommodate talker differences in speech production by adjusting phonetic boundaries in light of a talker’s unique productions. However, phonetic categories are not marked solely by boundaries, they also have a graded internal structure in that not all members of a category are considered equally good members. The current work examines whether listeners adjust the internal structure of a phonetic category for individual talkers and whether such adjustments would transfer to a novel word. Listeners participated in training and test phases. During training, two groups of listeners heard a talker produce “cane.” Voice-onset-time (VOT) was manipulated such that one group heard the talker produce /k/ with relatively short VOTs and the other group heard relatively longer VOTs. During test, listeners were presented with a VOT continuum from “gain” to “cane” or “goal” to “coal” and asked to rate each member for goodness as /k/. The results showed that the best exemplar region at test differed between the two groups in line with exposure during training, though this effect was greater for the training word compared to the novel word. These results suggest that internal category structure is dynamically adjusted in order to accommodate talker-specific phonetic variation.

Index Terms: speech perception, perceptual learning, talker-specificity

1. Introduction

A major goal of research in the domain of speech perception has been to describe low listeners achieve perceptual stability given that there is no one-to-one mapping between the acoustic signal and phonetic category members (i.e., consonants and vowels). Variability stems from many factors including speaking rate [1], phonetic context [2], and even idiosyncratic productions across individual talkers [3]. The latter source of variability, talker-specific phonetic variation, is the focus of the current work.

For many years, variability in the implementation of speech sounds across individual talkers was considered as problematic noise for the perceptual system. Traditional accounts of speech perception posited normalization mechanisms that served to remove variation in the acoustic signal, like that stemming from talker differences in pronunciation, in order to map the signal onto abstract or canonical representations [4]. However, there is now a growing body of evidence that challenges the existence of an information-reducing normalization mechanism.

At the heart of this evidence are findings indicating that experience with a particular talker’s voice facilitates subsequent linguistic processing. Listeners show faster processing times for familiar compared to unfamiliar talkers [5]. Familiar talkers are also more intelligible compared to unfamiliar talkers, particularly in noisy or degraded listening environments [6]. Findings from the memory literature suggest that listeners obtain these processing benefits, at least in part, by encoding in memory talker-specific phonetic variability [7]. Indeed, there is evidence that listeners store talker-specific phonetic instantiations and use this information to facilitate retrieval of individual lexical entries [8].

Given that listeners encode talker-specific phonetic detail and use it to facilitate linguistic processing, a complete theoretical account of speech perception must describe how such information is used to customize speech processing for individual talkers. To this end, recent findings suggest that listeners adjust for idiosyncratic productions at a prelexical level of representation; that is, at the earliest stages of mapping from the acoustic signal to meaning. For example, Norris and colleagues [9] demonstrated that when presented with a speech sound that is ambiguous between two categories, listeners use lexical information to modify the boundary between the two categories. Relevant to the current work, this lexically informed boundary adjustment is often applied on a talker-specific basis [10, 11, but see also 12]. Thus, one way that listeners accommodate a talker’s idiosyncratic productions, particularly for ambiguous productions, is to customize the initial mapping between the acoustic signal and speech segment by dynamically adjusting perceptual boundaries.

In addition to ambiguous productions, talker differences in phonetic properties of speech can also signal well-defined, unambiguous category members [13]. One such example is voice-onset-time (VOT), which marks the voicing contrast in English stop consonants. Even when contextual influences such as speaking rate are controlled, talkers differ in their characteristic VOTs, with some talkers having characteristic long VOTs than other talkers [14]. Moreover, listeners are sensitive to these differences. Theodore and Miller [15] exposed listeners to the speech of two female talkers. During training phases, characteristic VOTs were manipulated such that one talker produced /p/ with relatively short VOTs and the other talker produced /p/ with relatively longer VOTs. Following training, listeners were presented with a short-VOT and a long-VOT variant of /p/ and asked to select which was most representative of each talker. Their results showed that which variant was selected was contingent on exposure during training. Such data demonstrate that listeners can track VOT with respect to talker, with raises the possibility that listeners customize the mapping between the signal and phonetic category for individual talkers, as in the case of the ambiguous characteristic productions described above.

The current work tests this hypothesis by examining talker-specific contextual influences on the internal structure of phonetic categories. It has long been known that phonetic categories are not marked solely by boundaries; rather, they exhibit a rich, graded internal structure such that not all members of a category are considered equally good exemplars [16]. Moreover, the best exemplars of a category have been shown to shift as a function of context. For example, the best exemplars of the /p/ category are located at longer VOTs for a
that took approximately two hours to complete. Given that the present data were collected as part of a larger study to test listeners only using Joanne’s voice was methodological in present an equal number of voiced and voiceless tokens. Training set, the VOT tokens ranged in VOT from 78-87 ms, and VOTs for the long-VOT/short-VOT continua. Acoustically, successive steps on each continuum. To create the continua, a pitch-synchronous LPC analysis was performed on each token. These parameters were equated for word duration and root-mean-square (RMS) amplitude and each served as the voiced-initial endpoint of a continuum. To create the continua, a pitch-synchronous LPC analysis was performed on each token. These parameters were manipulated in order to synthesize additional tokens of the continua. Acoustically, successive steps on each continuum differed in VOT by approximately 4 milliseconds (ms).

From these tokens, two sets of training stimuli were created, one for each training group. All training sets consisted of the gain endpoint tokens, two tokens from the short-VOT cane region, and two tokens from the long-VOT cane region. For the J-SHORT/S-LONG training group, the short-VOT cane tokens were drawn from Joanne’s continuum and the long-VOT cane tokens were drawn from Sheila’s continuum. For the J-LONG/S-SHORT training group, the long-VOT cane tokens were drawn from Joanne’s continuum and the short-VOT cane tokens were drawn from Sheila’s continuum. Two amplitude variants of each selected token corresponding to RMS amplitude of the short-VOT and long-VOT tokens were created in order to eliminate a potential amplitude-based confound. In both sets of training stimuli, VOT for the gain tokens was 25 ms, the short-VOT cane tokens ranged in VOT from 78-87 ms, and VOTs for the long-VOT cane tokens range from 170-179 ms. Within each training set, the gain endpoint token was duplicated in order to present an equal number of voiced and voiceless tokens.

Test stimuli were the same for both training groups and consisted of tokens from Joanne’s continuum. The decision to test listeners only using Joanne’s voice was methodological in that the present data were collected as part of a larger study that took approximately two hours to complete. Given that previous work has shown that listeners can simultaneously track characteristic VOTs for two talkers (e.g., [15]), we chose here to test only on Joanne’s voice in order encourage feasibility of completion. The test stimuli consisted of 24 tokens that spanned VOTs from 25 ms to 183 ms. The step size of the first 12 tokens was approximately 4 ms and the step size of the last 12 tokens was approximately 8 ms.

To sum, training stimuli consisted of tokens of gain and cane produced by two talkers, with VOTs of the cane tokens manipulated across training groups. Test stimuli for both training groups consisted of Joanne’s gain to cane continuum.

2.1. Methods

2.1.1. Participants

Eighteen adults participated in this study. Half were assigned the J-SHORT/S-LONG training group and the other half were assigned to the J-LONG/S-SHORT training group. All were monolingual native speakers of American English who passed a pure-tone hearing screening on the day of testing and had no history of speech/language disorders.

2.1.2. Stimuli

Stimuli consisted of two synthesized continua ranging from gain to cane, each produced by one of two female talkers, fictitiously labeled “Joanne” and “Sheila.” Stimulus preparation followed the procedure outlined in Theodore and Miller [15]. To sum, naturally produced tokens of gain were obtained from the two female talkers. The tokens were equated for word duration and root-mean-square (RMS) amplitude and each served as the voiced-initial endpoint of a continuum. To create the continua, a pitch-synchronous LPC analysis was performed on each token. These parameters were manipulated in order to synthesize additional tokens of the continua. Acoustically, successive steps on each continuum differed in VOT by approximately 4 milliseconds (ms).

From these tokens, two sets of training stimuli were created, one for each training group. All training sets consisted of the gain endpoint tokens, two tokens from the short-VOT cane region, and two tokens from the long-VOT cane region. For the J-SHORT/S-LONG training group, the short-VOT cane tokens were drawn from Joanne’s continuum and the long-VOT cane tokens were drawn from Sheila’s continuum. For the J-LONG/S-SHORT training group, the long-VOT cane tokens were drawn from Joanne’s continuum and the short-VOT cane tokens were drawn from Sheila’s continuum. Two amplitude variants of each selected token corresponding to RMS amplitude of the short-VOT and long-VOT tokens were created in order to eliminate a potential amplitude-based confound. In both sets of training stimuli, VOT for the gain tokens was 25 ms, the short-VOT cane tokens ranged in VOT from 78-87 ms, and VOTs for the long-VOT cane tokens range from 170-179 ms. Within each training set, the gain endpoint token was duplicated in order to present an equal number of voiced and voiceless tokens.

Test stimuli were the same for both training groups and consisted of tokens from Joanne’s continuum. The decision to test listeners only using Joanne’s voice was methodological in that the present data were collected as part of a larger study that took approximately two hours to complete. Given that

2.2. Results

2.2.1. Training

Accuracy during the training phases was assessed separately for the phonetic decision and the talker decision. As shown in Figure 1, mean performance across listeners was near ceiling for both decisions, indicating that listeners were able to discriminate the talkers’ voices and that the synthesized tokens presented in training yielded the intended perceptual contrast.

Figure 1: Mean percent correct talker and phonetic decision for the talkers presented during training. Error bars indicate standard error of the mean.

2.2.2. Test

Recall that the test stimuli were identical for each training group and consisted of a gain to cane continuum produced by Joanne. Because our primary question concerns the degree to
which exposure during training guides performance at test, performance was analyzed separately for each training group. Mean goodness rating for each of the 24 test stimuli was calculated for each subject by collapsing across the six test phases. Figure 2 shows the mean goodness ratings as /k/ for each training group. Mean goodness ratings for both groups reflect previously established findings. Namely, short VOTs receive very low goodness ratings, because in fact these tokens are perceived as /g/. As VOTs increase, so do goodness ratings in line with VOTs becoming long enough to signal /k/. Critically, inspection of the figure suggests that the tokens rated as best exemplars for /k/ differed across the training groups. That is, the peak of the function for the J-SHORT/S-LONG training group is located at shorter VOTs than the peak of the function for the J-LONG/S-SHORT training group.

Figure 2: Group goodness functions as a function of VOTs of the “gain” to “cane” continuum for both training groups. The horizontal bars indicate the best exemplar range for each function.

To test the statistical significance of this pattern, performance was quantified using conventions outlined in Allen and Miller [18]. A best exemplar range was calculated for each subject, defined as the range of VOTs rated at or above 90% of the highest rating provided by that subject. The lower and upper bounds of the best exemplar regions was compared across the two training groups. The results showed that the lower bound of the best exemplar region was located at shorter VOTs for the J-SHORT/S-LONG training group compared to the J-LONG/S-SHORT training group [t(16) = -4.20, p = .006]. The same pattern was also observed for the upper bound of the best exemplar region [t(16) = -3.25, p = .005]. Thus, the peaks of the goodness functions were located at different VOTs across the training groups and, critically, in line with previous exposure to Joanne’s voice.

3. Experiment 2

The results from Experiment 1 suggest that sensitivity to a talker’s characteristic productions promotes a comprehensive re-mapping of acoustic phonetic space in line with that talker’s characteristic productions. However, a stricter test of such an account would entail assessing generalization to phonetic environments not presented during training. That is, the shift in internal category structure was assessed using the same word as presented during training (i.e., cane). Experiment 2 examines the nature of the shift in internal category structure by presenting a novel word during the test phase. If the shift in internal category structure observed in Experiment 1 reflects a talker-specific adjustment to the /k/ category, then we predict to observe the same shift in best exemplar range for the novel word tested in Experiment 2.

3.1. Methods

3.1.1. Participants

To date, 12 listeners not used in Experiment 1 participated in Experiment 2 following the previously outlined criteria.

3.1.2. Stimuli

The same training stimuli used in Experiment 1 were also used in Experiment 2. An additional set of test stimuli was created following the procedure outlined in Experiment 1. The test stimuli consisted of a goal to coal continuum produced by Joanne. The test set contained 24 tokens that were matched in VOT to the test tokens used in Experiment 1.

To sum, the same training stimuli used in Experiment 1 were used in Experiment 2 and consisted of synthesized tokens of gain and cane produced by two talkers, with characteristic VOTs of the cane tokens manipulated across training groups. Test stimuli were the same for both training groups and consisted of a goal to coal continuum produced by Joanne.

3.1.3. Procedure

Experiment 2 followed procedures outlined for Experiment 1.

3.2. Results

3.2.1. Training

Accuracy during training was analyzed as outlined for Experiment 1. As expected based on performance during training in the first experiment, mean performance across listeners was near ceiling for both the phonetic and talker decisions. These data are shown in Figure 3.

Figure 3: Mean percent correct talker and phonetic decision for the talkers presented during training. Error bars indicate standard error of the mean.

3.2.2. Test

Performance during test was assessed separately for each training group as outlined in Experiment 1. Figure 4 shows the mean goodness ratings as /k/ for each training group. Inspection of this figure suggests that the range of VOTs rated most prototypical of Joanne’s voice were again influenced by exposure during training. Namely, the peak of the J-SHORT/S-LONG function occurs at shorter VOTs compared
to the peak of the J-LONG/S-SHORT function. For each subject, a best exemplar range was quantified as outlined in Experiment 1 and the lower and upper bounds of this region were compared across training groups. Though numerically different, there was no statistically significant difference between the lower bound of the best exemplar range across the two training groups \(t(10) = -1.83, p = 0.09\). However, there was a reliable difference between the two training groups in terms of the upper bound of the best exemplar range, with it located at shorter VOTs for the J-SHORT/S-LONG training group compared to the J-LONG/S-SHORT training group \(t(10) = -3.80, p = 0.003\).

These results indicate that experience during training guided performance at test, even when tested on a word that was not presented during training. However, these data raise the possibility that such transfer of learning may not be as robust as for the word presented during training. This interpretation is tenuous given the smaller sample size of Experiment 2 compared to Experiment 1, and additional data should be collected prior to drawing firm conclusions.

![Figure 4: Group goodness functions as a function of VOTs of the “goal” to “coal” continuum for both training groups. The horizontal bars indicate the best exemplar range for each function.](image)

4. Discussion

Listeners are sensitive to a host of fine-grained phonetic variability, including that associated with individual talker differences in phonetic properties of speech. Listeners encode this variability in memory and use it to facilitate speech processing [7]. Many of the processing benefits associated with talker familiarity have been demonstrated at lexical levels of representation [e.g., 6]. However, there is a growing body of evidence indicating that the benefits observed at higher levels of processing may reflect adjustments listeners make at the earliest stages of speech perception [10].

In this vein, recent research has examined the nature of the perceptual adjustments listeners make in the initial mapping between the acoustic signal and linguistic representation. A consistent finding across many studies is that listeners are able to modify the boundary between phonetic categories in order to incorporate idiosyncratic productions, particularly when the production in question is ambiguous between two speech categories [e.g., 9]. However, as described in the Introduction, listeners are also sensitive to talker differences that result in well-defined, unambiguous category members. One such example is voice-onset-time specifying word-initial stop consonants; listeners can learn that one talker says /p/ with relatively short VOTs compared to a different talker’s productions, even though productions of both talkers clearly signal the intended stop consonant [15].

Given listeners’ sensitivity to talker differences within well-defined categorical space, the goal of the current work was to examine perceptual adjustments listeners make in order to accommodate these differences. The results suggest exposure to a talker’s prototypical productions results in a comprehensive re-mapping between the acoustic signal and phonetic category. Specifically, the category members considered the best exemplars of /k/ for the talker presented at test differed as a consequence of previous exposure to that talker’s voice. This effect does not appear to be contingent on the particular word presented during training, suggesting that listeners were making an adjustment to the phonetic category. However, additional data in the current Experiment 2 are necessary in order to confirm this possibility.

Collectively, the current work on talker-specific adjustments for clearly defined category members and previous work on talker-specific adjustments for ambiguous productions point to a perceptual system that dynamically adapts to a talker’s phonetic signature. When accommodating an ambiguous production, the categorical boundary space is adjusted. When accommodating clearly defined category members, the internal structure of the category is modified. Such findings demonstrate functional plasticity of speech perception that may result in processing advantages for higher levels of linguistic processing, including lexical retrieval.

One question that remains unanswered is the degree to which adjustments to phonetic boundaries versus internal category structure is contingent on the nature of a talker’s productions. That is, when accommodating productions that are ambiguous, is the perceptual adjustment limited to the boundary region such that the internal category structure does not shift? Likewise, does shifting internal category structure occur without a concomitant shift in the boundary region? Future research is aimed at examining this relationship.

5. Conclusions

Listeners are sensitive to a host of fine-grained phonetic detail, including that associated with individual talkers’ idiosyncratic productions [15]. Familiarity with a talker’s voice provides a host of processing advantages including faster [5] and more intelligible word recognition [6]. The results from the current work contribute to a host of findings indicating that listeners accommodate talker-specific phonetic detail at the earliest stages of speech perception, robustly modifying the mapping between the acoustic signal and individual speech segments on a talker-by-talkers basis.

6. Acknowledgements

Gratitude is extended to Janice Lomibao, Devin Roscillo, and Sarah Montanaro for assistance with stimulus creation and data collection. This research was supported by the College of Liberal Arts and Sciences, University of Connecticut, in the form of laboratory start-up funds to the author.
7. References


