Should I Stay or Should I Go? Talker-specific Influences on Distributional Learning for Speech

Nicholas Monto
University of Connecticut - Storrs, nicholas.monto@uconn.edu

Follow this and additional works at: https://opencommons.uconn.edu/dissertations

Recommended Citation
https://opencommons.uconn.edu/dissertations/2603
There is no one-to-one mapping between speech acoustics and individual speech sounds; instead, the acoustic cues produced for individual speech sounds show wide variability both within and across talkers. Nonetheless, listeners perceive the speech of familiar and novel talkers with ease. It is theorized that listeners achieve this feat by maintaining a degree of flexibility in how acoustics are mapped to speech sound categories, allowing listeners to dynamically modify the mapping to speech sounds to reflect structure in input statistics. Building on previous work demonstrating that listeners are sensitive to individual talker differences in speech production, we test the hypothesis that distributional learning for input statistics is contextually governed by talker identity. Listeners (n = 320) completed two blocks of phonetic identification for VOT input distributions specifying the /g/ and /k/ categories. In one block, the input was shifted towards relatively shorter VOTs; in the other block, the input was shifted towards relatively longer VOTs. Across listener groups, we (1) manipulated block order and (2) whether or not the talker remained constant across blocks. In this way, a change in input statistics was concomitant with a change in talker for some listeners but not for other listeners. Predictions for talker-specific vs. talker-agnostic distributional learning were derived through simulations performed with the Bayesian belief-updating model of speech adaptation, which yielded qualitatively different patterns of learning for the same-talker vs. different-talker simulations. Specifically, the simulations for the same-talker condition predicted that listeners in the two order groups would show a different VOT voicing boundary in Block 1 and then converge in their voicing
boundary in Block 2, consistent with a cumulative registration of input statistics for the same
talker heard across the two blocks. In contrast, the simulations for the different-talker condition
predicted that listeners in the two order groups would show a difference in the VOT voicing
boundary in both blocks, given a resetting of distributional learning in block two (i.e., a return to
prior knowledge) triggered by a change in talker. The results showed (1) robust evidence of
distributional learning in that listeners’ voicing boundaries moved block-to-block in line with
changes in the input statistics, (2) no difference between the same-talker and different-talker
conditions, and (3) learning patterns that were consistent with cumulative integration of
distributional input statistics across blocks. These patterns were replicated across two
experiments. Collectively, the results suggest that distributional learning in this paradigm is not
talker-specific, which may reflect the a priori informativity of VOT as a cue to talker identity.
Should I Stay or Should I Go? Talker-specific Influences on Distributional Learning for Speech

Nicholas R. Monto

B.S., University of Connecticut, 2014

A Dissertation
Submitted in Partial Fulfillment of the
Requirements for the Degree of
Doctor of Philosophy
at the
University of Connecticut
2020
Doctor of Philosophy Dissertation

Should I Stay or Should I Go? Talker-specific Influences on Distributional Learning for Speech

Presented by

Nicholas R. Monto, B.S.

Major Advisor

Rachel M. Theodore

Associate Advisor

Emily B. Myers

Associate Advisor

James S. Magnuson

Associate Advisor

Gerry T. M. Altmann

University of Connecticut

2020
Acknowledgments

I would like to thank my family and friends for putting up with me through the emotional highs and lows; my professors for imparting not only knowledge, but also the occasional nugget of wisdom; and my advisor, Dr. Rachel Theodore, for being patient and sticking with me for a bit longer than anticipated.
Table of Contents

1 Introduction
   1.1 Lack of invariance problem for speech perception 01
   1.2 Talkers provide structure to acoustic variability 03
   1.3 Listeners’ sensitivity to structured phonetic variation 07
   1.4 Talker-specific distributional learning in a Bayesian framework 13
   1.5 Computational simulations and predictions for the current work 16

2 Experiment 1
   2.1 Methods 23
   2.2 Results 27

3 Experiment 2
   3.1 Methods 33
   3.2 Results 33

4 Discussion 37

5 References 43

6 Appendix 47
1 Introduction

The current work examines the degree to which distributional learning of speech input statistics is linked to individual talkers. We set the stage by first discussing the lack-of-invariance problem for speech perception, which highlights the limited availability of explicit one-to-one mappings between speech sound cue patterns and speech sound categories. Next, we discuss evidence indicating that talker identity provides distinct structure to the variation in speech sound cues. Following this, we review evidence that listeners are sensitive to statistical structure in the acoustic-phonetic signal and possess the ability to dynamically modify the mapping to speech sounds given sufficient contextual information. We end the introduction by reviewing a computational model of speech adaptation that accounts for context-specific updating of speech sound cue distributions and presenting simulations performed with this model that form predictions for the current work.

1.1 Lack of invariance problem for speech perception

It is well established that the acoustic-phonetic cues listeners use to recognize speech sounds show considerable utterance-to-utterance variation due to a host of factors including speaking rate (Summerfield, 1981), phonetic context (Liberman, 1957), and even who in particular is speaking (Hillenbrand, Getty, Clark, & Wheeler, 1995; Newman, Clouse, & Burnham, 2001; Theodore, Miller, & DeSteno, 2009). Accordingly, successful language comprehension requires listeners to solve the lack of invariance problem, where in some cases, multiple acoustic forms are mapped to the same perceptual category, and in other cases, the same acoustic form can be mapped to different perceptual categories.

A study conducted by Hillenbrand and colleagues (1995) paints a picture with regards to the amount of variation that listeners encounter within American English vowels. In their study,
a host of acoustic cues including duration, F0, F1, F2, F3, and F4, were measured for productions of American English vowels. The acoustic cues were analyzed from a sample of 12 /hVd/ utterances spoken by 139 talkers (45 men, 48 women, 46 children). Hillenbrand and colleagues (1995) reported a large amount of variation across talkers and phonetic categories that could lead to the situation where, for example, the F1-F2 values produced for a given vowel by one talker might match the F1-F2 values produced by a different talker for a different vowel. Despite this large degree of overlap in the acoustic dimensions, listeners were still able to identify vowels with a high degree of accuracy (Hillenbrand et al., 1995). Perhaps by including additional information, not available in the discrete F1-F2 space, listeners were able to reduce the overlap across speech sound categories. Findings from Syrdal and Gopal (1986) suggest that listeners use F0 and F3 in order to scale F1-F2 space to improve recognition of vowels as the inclusion of F0 scaling helps account for variation across talker. However, additional features outside of the spectral domain might be used to further improve recognition. In fact, Hillenbrand and colleagues (1995) found that the inclusion of vowel duration led to moderate gains in classification accuracy, but including at least two samples of the chosen formant dimensions led to the largest gains in classification accuracy.

Past research has sought to identify areas of acoustic (e.g., acoustic landmarks; Stevens, 2002) and articulatory invariance (e.g., abstracted gestures in motor theory; Liberman & Mattingly, 1985) which listeners might leverage to elude the lack-of-invariance problem. Other research tackles speech perception by accounting for variability in speech sound cues, without attempting to minimize it, by identifying potential sources of structure within cues across speech sound categories that listeners may use to constrain the mapping to speech sound categories (e.g., Weatherholtz & Jaeger, 2016). This approach assumes that there are clear sources for the
structure present in the variability of low-level acoustic-phonetic cues, listeners are sensitive to fine-grain statistical structure, and that listeners are able to dynamically adapt to changes in fine-grain statistical structures. In the following sections, we discuss evidence to suggest that these assumptions are valid.

1.2 **Talkers provide structure to acoustic variability**

Vowel production data from both Peterson and Barney (1952) and Hillenbrand and colleagues (1995) show that variation in acoustic-phonetic information occurs at multiple levels. We observe that across gender and age, explicit speech sound categories can be formed using just the first two formants of the acoustic signal. Within these speech sound categories, average productions for men, women, and children produce three clearly distinct groupings that listeners may be able to leverage in rapid speech recognition. The distinctiveness of these groupings is primarily due to differences in the vocal tract, where men have typically longer vocal tracks than women and children, resulting in lower resonant frequencies. It has been shown that fundamental knowledge of articulators can be used as contextual information to help guide speech perception (Galantucci, Fowler, & Turvey, 2006; McGowan & Berger, 2009). This provides a precedent by which listeners might, to some degree, track variation in speech sound cues related to formant information in a talker-specific manner.

Systematic, talker-specific phonetic variation has also been observed for spectral cues that specify the /ʃ/-/s/ distinction (Newman, Clouse, & Burnham, 2001). Newman et al. (2001) studied the variability of centroid frequency and skew in word-initial /s/ and /ʃ/ productions across twenty talkers. They found that these cues (centroid frequency and skew) were strongly correlated across talkers such that talker-specific differences in speech production affected both cues (Newman et al., 2001). The authors also note the degree to which productions of /ʃ/ and /s/
overlapped across these dimensions varied drastically across talkers, such that some talkers produced very distinct categories and others produced much less distinctive categories (Newman et al., 2001). Of note, distributions of /ʃ/ and /s/ were more clearly formed along the centroid frequency dimension when focusing on within-talker variability as opposed to between-talker variability, providing evidence that talker can provide structure to phonetic variation that is observed across speakers of a language. How closely the speech sound categories were to each other in the defined cue space influenced listeners’ categorizations of the talker productions. Specifically, listeners were both faster and more accurate in categorizing productions that were further apart in the cue space compared to productions that were closer together (Newman et al., 2001). Here we see evidence that realizations of cue distributions vary in a talker-specific manner and that the degree to which distributions overlap for a given talker can influence speech perception.

Talker-specificity of statistical structures in cue distributions has also been found for cues outside of the spectral domain, including temporal cues that specify the voicing contrast for stop consonants (Allen, Miller, & DeSteno, 2003; Chodroff & Wilson, 2017; Theodore et al., 2009). Talker-specific differences have been observed in voice-onset-time (VOT) in both isolated and connected speech, particularly for voiceless stops (/p/, /t/, /k/) (Allen et al., 2003; Chodroff & Wilson, 2017; Theodore et al., 2009). Some of these differences may be linked to tertiary talker characteristics, such as speaking-rate (Allen et al., 2003; Chodroff & Wilson, 2017; Theodore et al., 2009). However, even when controlling for speaking rate, talker-specific differences in characteristic VOTs are present (Allen et al., 2003); some talkers have longer VOTs than others. The talker-specific influence of speaking rate on VOT has been found to be stable across a change in place of articulation, such that the magnitude of VOT displacement across voiceless
stop consonants, /p/ and /k/, did not vary for a given rate (Theodore et al., 2009). This would suggest that listeners would, having heard productions for one voiceless stop a given speaking rate, be able to extrapolate that knowledge to other voiceless productions at the same rate. Listeners do indeed display an acute sensitivity to characteristic VOT productions for a given talker, showing the ability to generalize to other productions at a specific speaking rate (Allen & Miller, 2004; Theodore & Miller, 2010; Theodore, Myers, & Lomibao, 2015). Outside of speaking rate, other contextual influences on VOT, including place of articulation, appear to reflect more language-general patterns (Chodroff & Wilson, 2017; Theodore et al., 2009). Together, the production and perception data to date suggest that listeners may be able to help solve the lack-of-invariance problem by learning a given talker’s idiosyncratic productions, which generalize across some contexts.

Generalized speech sound categories can be formed given information from one or two specific speech sound cues. These categories are able to then be navigated relative to other information available to the listener in order to achieve more specificity. Aside from talker-specific contexts, what other information might be leveraged by listeners to improve the specificity of generalized speech sound categories? Recent research has studied several degrees of socio-indexical groupings in order to help determine (1) which grouping might be most informative of underlying cue distributions and (2) which grouping might be most beneficial in phoneme recognition (Kleinschmidt, 2019). F1-F2 and VOT cue distributions were studied across several socio-indexical groups; gender, dialect, age, and talker. The four socio-indexical groupings represent varying degrees of information specificity; talker has the most specific information, followed by gender, age, and dialect. Marginal distributions were defined as the cue distributions averaged across all talkers, and thus contained the least specific information as to
the properties of any given speaker (Kleinschmidt, 2019). By comparing the cue distributions separated by each socio-indexical group to the marginal cue distribution using Kullback-Leibler divergence, Kleinschmidt (2019) obtained a metric of how different the grouped distributions were from the marginal, which was used to indicate a difference of informativity. The tested distributions were comprised of raw F1-F2 Hz values for American English vowels, Lobanov normalized F1-F2 values for American English vowels, and VOT values for American English stop consonants. The results showed that grouping by specific talker was found to be most informative of the cue distribution (i.e., most different from the marginal) for all cue distributions (Kleinschmidt, 2019). However, the degree of talker-specific informativity was drastically larger for spectral properties (i.e., F1-F2) compared to temporal properties (i.e., VOT). This difference is likely due to formants carrying specific information regarding the vocal tract structure of the talker (e.g., vocal tract length), while VOT is more influenced by properties outside of vocal tract structure (e.g., speaking rate, idiolect).

A second experiment looked at how well these socio-indexical groupings could be used to guide recognition of the vowel and stop categories. Category distributions were first estimated via the mean and variance of each phonetic category based on a subsample of the full dataset for each set in question: raw F1-F2, Lobanov normalized F1-F2, and VOT (Kleinschmidt, 2019). The derived speech sound categories of the subsample were tested against talkers that were either part of the subsample (talker-specific case); part of the same gender group, but part of the subsample; part of the same dialect group, but not the same subsample; or part of the same age group, but not the same subsample. Raw accuracy values were calculated by averaging across the posterior probabilities of a test talker’s intended categories. Log-odds were calculated to provide a metric of how much information is gained over random guessing (Kleinschmidt, 2019). As
before, information of the specific talker was found to have log-odds above and beyond information of the marginal distribution and other socio-indexical groups, in additional to having the highest raw accuracy in general (Kleinschmidt, 2019). This may be due to the fact that knowledge of the specific talker might contain a set of constraints surrounding articulatory information that is lost or muddled in the marginalized data and is not fully represented in gender-specific distributions. The difference of cue informativity may prove to influence the degree to which listeners leverage this information. The results of this study provide strong evidence that tracking cue distributions in a talker-specific manner provides the most benefit to categorical perception compared to tracking speech sound categories according to more general groupings.

We have presented evidence that there is a great deal of variation present in the acoustic-phonetic signal and that this variation appears to have systematic structure related to information present at multiple levels. While not explicitly necessary, we see that it may be beneficial for listeners to track cue distributions in a manner that is talker-specific, which may involve tracking multiple acoustic properties at the same time. This, however, makes the assumption that listeners are even sensitive to fine-grain phonetic detail and are able to track changes in a single cue, which we discuss further in the next section.

1.3 Listeners’ sensitivity to structured phonetic variation

As it so happens, listeners do indeed show a keen sensitivity to structured variability at multiple acoustic and linguistic levels (Clayards, Tanenhaus, Aslin, & Jacobs, 2008; Kleinschmidt & Jaeger, 2015; Romberg & Saffran, 2010; Schriber, Onishi, & Clayards, 2013; Theodore & Monto, 2019). One statistical measure that listeners appear to be sensitive to are transitional probabilities at varying levels of granularity (e.g., Aslin, Saffran, & Newport, 1998).
These translational probabilities are generally related to conditional probabilities between different categories, which can be described via a simple mathematical formula, Bayes’ theorem (Bayes & Price, 1763), as shown in equation (1).

\[
p(A \mid B) = \frac{p(B \mid A)p(A)}{p(B)} \quad (1)
\]

This theorem states that the probability of the occurrence of A given B is equal to the probability of B given A multiplied by the independent probability of A, over the independent probability of B. A simplified version of this formula has been applied to describe listener sensitivity to distributional information in the acoustic-phonetic signal.

With respect to acoustic-phonetic input statistics, Clayards et al. (2008) used the ideal observer computational framework to examine listeners’ sensitivity to distributional information for speech perception. The exact model is shown in equation (2).

\[
p(\text{Category A} \mid \text{Stimulus X}) = \frac{p(\text{Stimulus X} \mid \text{Category A})}{p(\text{Stimulus X} \mid \text{Category A}) + p(\text{Stimulus X} \mid \text{Category B})} \quad (2)
\]

This model provides an explicit framework by which listeners would be able to track statistical regularity present in the acoustic-phonetic signal. Presuming the ability to track fine-grain acoustic-phonetic detail, perception occurs as a result of comparing the probability of perceiving an acoustic-phonetic cue (e.g., VOT) given the speaker’s intended categorical production (e.g., voiced or voiceless stop consonant) against an opposing category. These probabilities reflect the distributional information in the input. Clayards and colleagues (2008) presented two groups of listeners with VOT distributions specifying /b/ and /p/ categories. One group of listeners heard
VOTs that yielded minimal variance for the two categories (the narrow input). The other group heard VOTs that yielded maximal variance for the two categories (the wide input). The means of the /b/ and /p/ categories were identical between the narrow and wide listener groups (Clayards et al., 2008). Their results showed that listeners were sensitive to the underlying statistics; listeners presented with narrow distributions responded with more certainty, reflected by a steeper slope of the identification curve, compared to listeners who were presented with wide input distributions, who showed a shallower identification curve (Clayards et al., 2008).

Evidence of sensitivity to structured phonetic variation has led to the development of theories concerning perceptual learning, which posit that listeners track phonetic cues in the input with respect to a higher-order structure, such as talker identity, and use this information to derive a structure-specific probabilistic mapping that optimizes phonetic categorization (e.g., McMurray, Aslin, & Tuscano, 2009; Kleinschmidt & Jaeger, 2015).

Efficient speech perception involves tracking of multiple acoustic dimensions of the speech signal. However, the degree to which listeners use each acoustic dimension when perceiving a given speech sound varies. Examples of this can be found in both temporal and spectral speech sound cues. Within stop consonants, for example, there lies a relationship between the initial stop consonant and the F₀ of the following vowel such that when the initial stop consonant is voiced, vowel onset F₀ tends to be lower compared to vowel onset F₀ following a voiceless stop consonant (Idemaru & Holt, 2011; Idemaru & Holt, 2014; Idemaru & Holt, 2020). Within the vowels /ɛ/ and /æ/, a relationship exists between steady-state formant frequency (i.e., spectral quality) and vowel duration, such that lower spectral quality is associated with short vowel duration and higher spectral quality is associated with relatively longer vowel duration (Liu & Holt, 2015). Typically, one of the cue dimensions is favored over
the other. For stop consonant voicing, VOT is the dominant cue (Idemaru & Holt, 2011; Idemaru & Holt, 2014; Idemaru & Holt, 2020), while formant frequency is the dominant cue for vowels (Liu & Holt, 2015). When acoustic cue relationships shift (e.g., low VOT- high vowel onset $F_0$), it is perceived as an accent (Idemaru & Holt, 2011; Idemaru & Holt, 2014; Idemaru & Holt, 2020). Listeners have been found to dynamically shift their perceptual weighting of these cue dimensions given exposure to token distributions that violate standard cue weightings (Idemaru & Holt, 2011; Idemaru & Holt, 2014; Idemaru & Holt, 2020; Liu & Holt, 2015). For the $F_0$-VOT relationship, when exposed to token distributions reflecting the canonical cue relationship, listener respond to test tokens with an ambiguous VOT and low vowel onset $F_0$ as more voiced compared to tokens with an ambiguous VOT and high vowel onset $F_0$ (Idemaru & Holt, 2011; Idemaru & Holt, 2014; Idemaru & Holt, 2020). However, when exposed to token distributions representing a reversed $F_0$-VOT relationship, listener responses to ambiguous test tokens indicate a perceptual down-weighting of vowel onset $F_0$, such that there were more voiceless responses for test tokens with a low vowel onset $F_0$ in reversed blocks compared to canonical blocks (Idemaru & Holt, 2011; Idemaru & Holt, 2014; Idemaru & Holt, 2020). Perceptual down-weighting has been observed to generalize to novel word forms (Idemaru & Holt, 2020) and productions from a novel talker (Liu & Holt, 2015). However, the generalization is never to the same degree as observed in previously exposed instances (Idemaru & Holt, 2020; Liu & Holt, 2015), suggesting that listeners cumulatively integrate short-term exposure with prior knowledge formed over long-term experience.

Thus far, we have reviewed evidence indicating that listeners display both a sensitivity to statistics present in the acoustic-phonetic signal (e.g., Clayards et al., 2008) as well as the ability to dynamically adapt to changes in how cues are used by talkers (e.g., Idemaru & Holt, 2020). It
is important to note that listeners learn the statistical relationships of cues over extended exposure given experience with a language. Over time, the canonical relationships described previously become expected, and, in a Bayesian framework, serve as prior knowledge that listeners can draw from when encountering new input. When listeners encounter deviations from expected speech input, such as the “reversed” relationship between VOT and vowel onset F₀ described above, listeners adapt by down-weighting their use of the less informative cue in the current speech input, the cue which is deviant from their expectations (Idemaru & Holt, 2011; Idemaru & Holt, 2014; Idemaru & Holt, 2020; Liu & Holt, 2015). However, even in the case of days of repeated exposure to the reversed cue relationship, listeners stubbornly hold on to their prior knowledge of the long-exposed canonical cue relationship (Idemaru & Holt, 2011).

Part of the hesitancy to fully adapt to the novel reversed cue relationship may be due to its extreme deviance from the expected canonical relationship. Hesitance to adapt to the presence of novel statistical structure has been observed in research focused on transitional probabilities of syllabic and novel lexical structures, also called the primacy effect (Bulgarelli & Weiss, 2016; Gebhart, Aslin, & Newport, 2009; Zinszer & Weiss, 2013). It is posited that these primacy effects are a result of “overlearning” the initial structure and reduced attention to future regularities (Bulgarelli & Weiss, 2016). This overlearning can also be described as hyper-confidence, in which the system needs to be presented many items that greatly contradict the prior, or be reset through the use of some explicit, external cue. Bulgarelli and Weiss (2016) found that entrenchment could be overcome by priming the system with an explicit cue; namely, telling listeners that they will be presented items from multiple languages. Zinszer and Weiss (2013) have likewise observed a reduction in entrenchment through the presentation of multiple structures with unique statistical regularities. By interleaving multiple statistically distinct
structures during exposure, the perceptual system is kept “on its toes,” so to speak, minimizing the chance of the system to become over confident in one source of regularity over another.

A reduction of entrenchment has also been found to occur when providing listeners with an additional cue to the presence of a new statistical structure, like linking it to a change in talker (Weiss, Gerfen, & Mitchel, 2009). While this may work for syllable-level statistical regularities, it is unclear whether listeners learn and maintain statistical regularities for phonetic distributions across talkers. There is behavioral evidence suggesting that when provided with acoustic-phonetic experience with a specific talker, listeners often receive processing benefits (Clarke & Garrett, 2004; Nygaard, Sommers, & Pisoni, 1994). However, the explicit mechanism behind these processing benefits is unknown, though it appears that these benefits reflect adjustments that listeners make early in the processing stream (e.g., Bradlow & Bent, 2008; Eisner & McQueen, 2005; Kraljic & Samuel, 2005).

There is clear evidence that listeners are able to track and adapt to statistical variation at multiple levels of the speech signal. There are cases, however, in which adaptation is not observed, like attempting to learn structures with competing statistical regularities (Weiss, Gerfen, & Mitchel, 2009). In these cases, it listeners appear to be able to learn the competing regularities by linking them to some additional cue, like talker (Weiss, Gerfen, & Mitchel, 2009) or language (Bulgarelli & Weiss, 2016), allowing listeners to adapt to regularity in a different space. It seems that listeners are able to reconcile overlapping statistical signals by tying them to differential contexts. It would be useful to be able to capture these effects in a singular computational model that would allow researchers to generate testable predictions as to how listeners will respond to variable distributional input in the acoustic-phonetic signal given different contexts. The addition of context is important in adapting to overlapping cue
distributions, which occur to a staggering degree is acoustic-phonetic cues. While Bayes rule has promise for predicting behavioral responses to acoustic-phonetic cues within a single dimension (e.g., Clayards et al., 2008), it is not, in its basic iteration, able to account for potential tracking of overlapped cue distributions across multiple groupings, such as talker and gender. The next describes a computational model of speech adaptation that can in fact address these concerns.

1.4  *Talker-specific distributional learning in a Bayesian framework*

The Bayesian belief-updating model of speech adaptation (Kleinschmidt & Jaeger, 2015) is a model of speech perception that makes the assumption that speech productions are random samples from distributions corresponding to relevant speech sounds. Perception is the result of listeners attempting to estimate these generative distributions by integrating new information in an iterative fashion (Kleinschmidt & Jaeger, 2015). The key algorithm for this model is shown below in equation (3).

\[
p(\mu_c, \sigma_c^2|X, C = c) \propto p(X|\mu_c, \sigma_c^2, C = c)p(\mu_c, \sigma_c^2)
\]

This formula is very similar to Bayes’ theorem, but places specific constraints on the underlying distributions, in this case the distributional information is assumed to be Gaussian and specific to context C. The expectations of the system are captured by the presence of a prior, \(p(\mu_c, \sigma_c^2)\). Prediction occurs via the process of comparing the sample of the environment (i.e., a specific observation), to the expectation of the system, via sampling of the prior. Adaptation, in this framework, occurs via the modification of the prior in light of observations in the environment.

Theodore and Monto (2019) tested predictions made by two Bayesian frameworks of distributional learning. In one framework, they used a modified Bayes rule that does not take into
account prior experience (equation 2). In the other framework, we used Kleinschmidt & Jaeger’s (2015) Bayesian belief-updating model, which takes into account prior experience and includes parameters for other sources of contextual information. We found that online identification reflected the cumulative integration of input statistics with prior knowledge (Theodore & Monto, 2019). In our study, listeners completed two blocks of phonetic categorization for stimuli that differed in VOT; the same speaker produced the tokens in each block. These stimuli were drawn from two continua (goal–coal, gain–cane). In each block, two distributions were presented, one specifying /g/ and one specifying /k/. Across the two blocks, the variance of the distributions was manipulated to be either narrow or wide, and we manipulated block order across two groups of listeners between two listener groups (Narrow-Wide vs. Wide-Narrow). The primary dependent measure was the slope of the identification function in each block. Our results showed that listeners tracked the VOT distributions in a cumulative manner such that the slope of the identification function was steeper for the Narrow-Wide group compared to the Wide-Narrow group in block one, when the statistical experience with the talker’s distributions differed between the two groups, but the two groups showed no difference in the identification slope in block two, when statistical experience between the two groups was equivalent (given that listeners had now heard both input blocks). This was in line with predictions of cumulative learning made via the ideal observer model. An interesting observation from this study was that convergence between the two order groups in Block 2 reflected by-block movement of only one order group; the slope of the identification function changed from Block 1 to Block 2 for the Narrow-Wide group, but remained constant across block for the Wide-Narrow group.

Theodore and Monto (2019) replicated the behavioral asymmetry in learning as a function of block order through computational simulations using Kleinschmidt and Jaeger’s
Bayesian belief-updating model, which suggests that this pattern of learning reflected the cumulative integration of exposure in each block with prior distributional knowledge. The asymmetry in this case was explained by the degree of overlap between the narrow and wide distributions. When the wide distributions were presented after the narrow distributions, new information is presented to the system in that the VOTs of the wide distributions fall outside the range of VOTs in the narrow distribution. But in the other case, where narrow distributions are presented after the wide distributions, no new information is presented because the narrow distributions are contained within the wide VOT distributions, thus the system did not need to update beliefs in the distributions to the same degree. The point of convergence did not change across order groups; instead; it was the trajectory of adaptation that changed as a function of order. This explanation converges with theories of entrenchment for statistical regularity. Specifically, in order for the system to markedly adapt to statistical regularities, these regularities must be different enough from previously learned regularities.

Thus far, we have reviewed evidence indicating that listeners possess a keen sensitivity to statistical information present in the acoustic-phonetic signal (Clayards et al., 2008; Idemaru & Holt, 2020; Theodore & Monto, 2019; Weatherholtz & Jaeger, 2016). Listeners are able to adapt to changes in the systematic variability within a talker (Idemaru & Holt, 2020; Theodore & Monto, 2019). Observable changes in behavioral responses only occur given sufficient difference between presented statistical structures or when linking statistical structures to other cues (Bulgarelli & Weiss, 2016; Weiss, Gerfen, & Mitchel, 2009; Zinszer & Weiss, 2013). Other research has shown that listeners are able to generalize previously learned changes in statistical cue relationships to a novel talker given sufficient spectral overlap (Liu & Holt, 2015). Within the distributional learning framework, adaptation to information present in the acoustic-phonetic
signal occurs at different contextual levels, depending on which would be most beneficial for the listener (Kleinschmidt & Jaeger, 2015; Kleinschmidt, 2019; Weatherholtz & Jaeger, 2016). Listeners may adapt to variation at the level of the marginalized speech sound category, the gender-specific speech sound category, or the talker-specific speech sound category. The degree of adaptation within each level might vary as a function of the environment. For example, if a listener is new to a region and expects to encounter many different talkers across a given span of time, adapting to changes at a marginalized speech sound level may be most advantageous. However, if a listener only expects to only hear one or two talkers within a short time span, perhaps it would be most advantageous to adapt their speech sound categories at the level of gender or talker. In the current work, we test the hypothesis that listeners track distributional cues in speech input with respect to a given talker, leading to talker-specific patterns of distributional learning. In the following section, we describe the goals of the current study in detail and present a series of computational simulations within the Bayesian belief-updating framework that generated testable predictions for the current work.

1.5 Computational simulations and predictions for the current work

Given previous evidence of listeners being able to differentiate and learn signals that overlap in a particular cue by linking them to specific talkers, we predict that listeners will show talker-specificity in how they adapt to changes in input statistics. Past work, including our own, has come to this conclusion without explicitly testing this possibility (e.g., Clayards et al., 2008; Theodore & Monto, 2019; Saltzman & Myers, 2018). For example, in Theodore and Monto (2019), we concluded that listeners tracked a talker’s input statistics cumulatively over time. However, listeners in that study only ever heard one talker; thus, it is not clear whether listeners were retuning the mapping to speech sound categories for that talker specifically, or rather in a
more general way that would be applied to a novel talker. In the current work, we provide a stricter test of the talker-specificity hypothesis by specifically manipulating whether listeners hear input from only one or from two talkers over time.

To preview our methods, described in detail below, listeners completed two blocks of phonetic identification for VOT input distributions that specified the /g/ and /k/ categories. In one block, VOTs were shifted towards shorter values; in the other block, VOTs were shifted towards longer values. The specific VOT input in each block is shown in Figure 1 (and Table 1). Across listener groups, we manipulated (1) the order in which each block was encountered and (2) whether the talker remained constant across blocks (same talker) or not (different talker). With this design, the change in distributional input across blocks was concomitant with a change in talker for the different talker conditions, but this was not the case for the same talker conditions. Two experiments were conducted using this design; the only difference between the two experiments was which talker was assigned to the same talker condition; thus, the two experiments serve as replications of each other.

Predictions for talker-specificity in adaptation were generated through computational simulations using the Bayesian belief-updating model of speech adaptation proposed by Kleinschmidt and Jaeger (2015). This model allows for phonetic cue distributions to be learned and updated as a function of a specific context (e.g., talker), which is accomplished by the resetting of prior knowledge when a change in talker is encountered. Specifically, this model posits that when hearing a novel talker, listeners make perceptual decisions based on prior expectations, formed given experience with a language. If the input deviates from expectations based on prior knowledge, then learning will occur iteratively (i.e., observation-by-observation) instantiated as updated priors that reflect an integration of the new evidence with the existing
priors. This iterative updating continues to occur until a listener is presented with a new talker, at which point priors are reset to the initial expectation (based on experience with a language in general), and then iterative updating of prior knowledge occurs again given evidence in the new context. While this has been proposed theoretically, it has been yet to be explicitly tested. Here we use the Bayesian belief-updating model (Kleinschmidt, 2017) and a helper package, slopeExtractR (Monto, 2018), to simulate the behavioral responses of listeners in the current study following the theory that listeners will cumulatively integrate observed evidence with prior knowledge, conditioned on talker as context for cumulative integration. These simulations were then used to derive predictions for listeners tested in the current work.

**Figure 1.** Panel A shows the prior distributions used for the computational simulations presented in the main text. Panel B shows the probability distributions for the short-VOT (Short input) and long-VOT (Long input) blocks. Panel C shows the mean predicted category boundary for the same talker and different talker conditions in each block for each order group. Means were calculated over the 40 simulated listeners in each talker/order grouping; error bars indicate standard error of the mean.
One hundred and sixty simulations were run to simulate 40 listeners in each of four conditions formed by crossing two levels of input order (Short-Long vs. Long-Short) and two levels of talker variability (same talker vs. different talker). The belief_update() function of the beliefupdatr package was used to iteratively update a specified normal $\chi^{-2}$ prior for each of two perceptual categories, /g/ and /k/, with the specified hyper-parameters of mean, variance, and confidence. Trial-by-trial observations of the perceptual parameter (e.g., VOT) and the response category (e.g., /g/ or /k/) were provided as input to the Bayesian algorithm. With this input, the learning algorithm (Kleinschmidt & Jaeger, 2015) updated the category-specific distributions on each trial by integrating the observed VOT and response with the prior distribution, weighted by confidence. The posterior distributions obtained after each trial reflected the likelihood of the prior distribution given the observed evidence. These posteriors were used to generate identification functions from which the category boundary was derived. The code to reproduce all simulations reported in this manuscript is available at: https://osf.io/wn34y/?view_only=0a216205aad3491d9175cfecaf10ebb69.

For each talker condition, we simulated 80 lists specifying trial-level VOT presentation for 304 trials. Each list thus provided a simulation of a single participant. Forty lists simulated trial-level VOT presentation for the Short-Long group; the first 152 trials were a unique randomization of VOTs presented during the short block and the second 152 trials were a unique randomization of VOTs presented during the long block. The number of each token and associated VOT is shown in Table 1 and visually presented in Figure 1B. The other forty lists simulated trial-level VOTs for the Long-Short order group, which followed the same procedure outlined for the Short-Long order group save that for the order in which the short and long blocks were presented. Response patterns for the 80 simulated listeners in each talker condition
matched the intended category for all VOTs.

Nine simulations (representing three prior specifications crossed with three confidence specifications) were performed for these lists in each of the same talker and different talker conditions, For all prior specifications, the standard deviation of priors was set to match that of a “typical talker” (SD = 8.3 for /g/ and 18.9 for /k/, Kleinschmidt & Jaeger, 2016) as presented in Figure 1A. Across the three prior specifications, means for /g/ and /k/ were set to be consistent with those present in the short block (/g/ = 32 ms, /k/ = 80 ms), shifted down 10 ms (/g/ = 22 ms, /k/ = 70 ms), or shifted up 10 ms (/g/ = 42 ms, /k/ = 90 ms). All of these prior specifications are reasonable given existing speech production data (e.g., Theodore et al., 2009; Chodroff & Wilson, 2017; Chodroff & Wilson, 2018). For each prior specification, confidence was set to 50, 100, and 200, values that represent relatively less to relatively more confidence in the prior specification, respectively, and spanning the range of inferred confidence reported previously (Kleinschmidt & Jaeger, 2016). The qualitative patterns held across all prior specifications and confidence levels. For brevity, we present results for the prior specification matched to input in the short block and confidence level of 50 in the main text, and show results of all simulations in the Appendix.

Simulations for the same talker vs. different talker conditions were implemented by simulating cumulative updating to priors throughout the 304 observations in each list (i.e., 152 trials in each of the short and long input blocks) for the same talker conditions and resetting priors to the initial state at observation 153 for the different talker conditions (i.e., resetting priors at the start of the second block, which in our design is concomitant with a change in talker for the different talker conditions). Put another way, in the same talker conditions, the initial priors were cumulatively updated as all 304 observations were presented to the model (consisting of either
152 observations from the short input followed by 152 observations from the long VOT input, or the same observations but in the opposite order). However, for the different talker condition, the initial priors were cumulatively updated through the first 152 observations (i.e., the end of the initial block of input), reset to the initial priors at trial 153, and then cumulatively updated through the second 152 observations.

In all simulations, category boundaries were derived for each simulated participant based on the inferred posterior distribution (i.e., updated prior) at trials 114 and 266 (38 trials from the end of the first and second blocks, respectively) by first calculating the identification function for the inferred posterior distributions at these trials and then extracting the VOT corresponding to 50% /k/ responses. These trials were selected to reflect a point in time at which posterior distributions would sufficiently reflect exposure in each block while still allowing the ability to observe possible effects of the specific observation orderings across the 40 simulated listeners in each condition that are extinguished by the last trial in each block, noting that the qualitative patterns we described below were incredibly consistent regardless of which specific trials near the end of each block were used. Figure 1C shows the derived category boundaries for listeners in the same talker and different talker conditions, respectively.

Consider first the average derived category boundaries for the same talker condition. With respect to between-subjects comparisons, the category boundary differs between the two order groups in block one, but converges between the two order groups in block two. This reflects different experience in the first block between the two order groups, and shared cumulative experience between the two order groups at the end of block two. Because the same talker simulations reflect cumulative updating of prior knowledge across blocks, the two groups converge in their beliefs at the end of the exposure period. With respect to within-subjects
comparisons, the simulations predict a large movement in the posterior boundary over time for the Short-Long order group, moving from a relatively shorter boundary to a relatively longer one over time, in line with the block-specific input distributions. However, there is minimal boundary movement observed in the Long-Short group, where the boundary remains relatively consistent despite the change in distributional input. This may reflect a point of convergence that is reached when the distributional information differs sufficiently from the prior.

Now consider the average category boundaries derived for the different talker condition. A robust between-subjects difference is observed in each block; specifically, the derived category boundary for the Short-Long order compared to the Long-Short order is at a shorter VOT in block one and a longer VOT in block two. Thus, in contrast to the same talker condition, the model simulations show a difference in the boundary between the two order groups in each block, reflecting block-specific input distributions. In terms of the within-subjects predictions, the simulations show a dynamic shift in the category boundaries across blocks for both order groups. This pattern reflects the fact that the model priors for these simulations were reset to the initial prior at the start of the second block in order to instantiation the theoretical claim that distributional cues are tracked with respect to individual talkers. (Kleinschmidt & Jaeger, 2015).

These simulations lead to clear qualitative predictions for the current work. If listeners modify the mapping to speech sounds to reflect talker-specific experience with distributional cues, then listeners in the two order groups for the same talker condition should show a difference in their perceptual responses in block one, reflecting initial experience that differs between the two order groups, and a convergence in their perceptual responses in block two, reflecting shared cumulative experience. Moreover, listeners in the two order groups in the different talker conditions will show divergence in their perceptual responses in both blocks, in
line with the input distributions that differ across blocks in terms of both statistical cues and talker identity. A failure to observe these patterns would suggest that distributional learning for VOT input cues does not reflect talker-specific learning.

2 Experiment 1

2.1 Methods

Participants. Participants were recruited from the Prolific participant pool (www.prolific.co). The participants (n = 160; 81 women, 79 men)\(^1\) were monolingual speakers of American English between 18 and 35 years of age (mean = 27, SD = 5) with no history of language-related disorders who were currently residing in the United States. Participants were randomly assigned to one of four between-subjects groups (n = 40 in each group) formed by crossing two levels of condition (same talker vs. different talker) and order (short-long vs. long-short). All participants provided informed consent following procedures approved by the University of Connecticut Institutional Review Board.

Stimuli. The stimuli consisted of voice-onset-time (VOT) continua that perceptually ranged from goal to coal, one for each of two female talkers (TD and TY) who had perceptually distinct voices. Each continuum consisted of 12 tokens; across continuum steps, VOTs ranged from 18 – 118 ms in approximately 10 ms steps. The continua were based on a subset of tokens used in Theodore and Miller (2010), which were created using an LPC-based speech synthesizer to successively increase word-initial VOT of a natural production of goal by changing successive voiced frames to voiceless frames following procedures outlined in Allen and Miller (2004). In order to create the two continua for the current study, the Praat Vocal Toolkit (Corretge, 2019)

\(^1\) An additional 41 participants were tested but excluded due to failure to pass the headphone screen, failure to exhibit a logistic response function in one or both of the test blocks, and/or showing a category boundary exceeding 40 ms of the intended category boundary in each block.
was used to manipulate $F_0$ of the base continuum (from stimuli used in Theodore and Miller, 2010) to instantiate a 45 Hz difference between mean $F_0$ for the two talkers. Specifically, talker TD had a mean $F_0$ of 229 Hz and talker TY had a median $F_0$ of 184 Hz. This procedure was used in order to equate tokens between the two talkers in every aspect except for talker identity. All tokens were equated for amplitude using Praat (Boersma & Weenik, 2020).

The stimuli used here (12 tokens x 2 talkers) were pre-tested in order to ensure that the two continua were indeed perceived as being produced by different talkers while also eliciting equivalent VOT response functions. In the pre-test, listeners ($n = 80$, consisting of the same demographic characteristics described for the main experiments) completed a talker discrimination task and a phonetic identification task. During talker discrimination, participants were presented with a pair of tokens that were produced by either the same talker or by different talkers. VOT was matched in a given pair; across pairs, VOTs were sampled across the continuum range (i.e., 18, 32, 98, and 118 ms). The discrimination task consisted of 96 trials, reflecting 48 “same” pairs and 48 “different” pairs; same pairs were evenly distributed between talkers, as was the order in which each voice appeared for different talker trials. On each trial, participants were asked to identify whether the two tokens were produced by the same talker or by different talkers; no feedback was provided. During phonetic categorization, participants completed two blocks of phonetic categorization, one for each talker. In each block, stimuli were presented in eight cycles, with each cycle consisting of a different randomization of the 12 continuum steps for the respective talker. On each trial, participants were asked identity whether each token began with either the /g/ or /k/ sound. No feedback was provided, and talker order was counterbalanced across participants.

Figure 2 shows the distribution of mean accuracy (proportion correct) across participants.
for the talker discrimination task (panel A), the distribution of mean sensitivity (d’) across participants for the talker discrimination task (panel B), and mean proportion /k/ responses as a function of VOT in the phonetic identification task (panel C). Results from the talker discrimination task show high accuracy (mean = 0.93 ± 0.07) and high sensitivity (mean = 3.37 ± 0.85), indicating that our stimulus creation methods successfully yielded stimuli that were perceived as two distinct voices. Trial-level data for all experiments presented in this manuscript, including the stimulus pre-test, and an analysis script that will reproduce all test statistics reported in this manuscript (in addition to generating all figures) is available at:

https://osf.io/wn34y/?view_only=0a216205aad3491d9175cfeca10ebb69.

**Figure 2.** Results of the stimulus pre-test. Panel A shows the distribution of talker discrimination accuracy across participants. Panel B shows the distribution of talker discrimination sensitivity (d’) across participants. Panel C shows mean proportion /k/ responses as a function of VOT for each talker; means reflect grand means calculated over by-subject averages; error bars indicate standard error of the mean.

Visual inspection of the results from the phonetic identification task show comparable response functions for the two talkers. To confirm this pattern statistically, trial-level responses (0 = /g/, 1 = /k/) were fit to a generalized linear mixed-effects model (GLMM) with the binomial response family using the glmer() function from the lme4 package (Bates, Maechler, Bolker, &
Walker, 2015) in R. The model contained fixed effects of VOT, talker, and their interaction. VOT was entered into the model as a continuous variable (scaled/centered around the mean); talker was contrast-coded (TD = -0.5, TY = 0.5). The model also included random intercepts by subject, and random slopes by subject for VOT, talker, and their interaction. The results of the model showed a main effect of VOT ($\hat{\beta} = 4.126, SE = 0.150, z = 27.600, p < 0.001$), no main effect of talker ($\hat{\beta} = -0.140, SE = 0.154, z = -0.910, p = 0.363$), and no interaction between talker and VOT ($\hat{\beta} = 0.024, SE = 0.216, z = 0.110, p = 0.913$). Collectively, the results of the stimulus pre-test confirm that the stimuli are suitable for the current work: (1) as indexed by the high d’ scores, tokens within each continuum were identified as being produced by the same talker and tokens across continua were identified as being produced by different talkers, and (2) as indexed by phonetic identification responses, the two continua elicited equivalent expected VOT response functions.

For the primary experiments, the tokens for each talker were arranged into sets that formed either short-VOT or long-VOT input distributions as shown in Table 1 (and Figure 1B). Each set contained 152 tokens. In the short-VOT input distributions, mean VOT was 34 ms (SD = 11 ms) for the /g/ category and 81 ms (SD = 11 ms) for the /k/ category. In the long-VOT input distributions, mean VOT was 54 ms (SD = 13 ms) for the /g/ category and 99 ms (SD = 11 ms) for the /k/ category.

Table 1. Frequency (number of tokens) of each VOT in the short-VOT input block (Short) and the long-VOT input block (Long) for the computational simulations and primary experiments.

<table>
<thead>
<tr>
<th>VOT (ms)</th>
<th>18</th>
<th>28</th>
<th>32</th>
<th>42</th>
<th>56</th>
<th>66</th>
<th>72</th>
<th>80</th>
<th>92</th>
<th>98</th>
<th>108</th>
<th>118</th>
</tr>
</thead>
<tbody>
<tr>
<td>Block</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Short</td>
<td>10</td>
<td>16</td>
<td>24</td>
<td>16</td>
<td>10</td>
<td>10</td>
<td>16</td>
<td>24</td>
<td>16</td>
<td>10</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Long</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>16</td>
<td>24</td>
<td>16</td>
<td>10</td>
<td>16</td>
<td>24</td>
<td>16</td>
<td>10</td>
<td></td>
</tr>
</tbody>
</table>
**Procedure.** All testing took place online via Gorilla Experiment Builder (www.gorilla.sc; Anwyl-Irvine, Massonnié, Flitton, Kirkham, & Evershed, 2020). Participants first completed the Woods et al. (2017) headphone screen designed to measure compliance with headphone use in web-based studies. Following the headphone screen, participants completed two blocks of phonetic categorization, one for the short-VOT input and one for the long-VOT input, with block order determined by assignment to experimental groups (e.g., listeners in the Short-Long order groups completed the short-VOT input block followed by the long-VOT input block). Listeners in the same talker groups heard the same talker (TD) in both blocks. Listeners in the different talker groups heard talker TY in the short-VOT input block and talker TD in the long-VOT input block.

In each block, participants were presented with one randomization of the 152 tokens that formed either short or long input distributions, as outlined above. On each trial, participants indicated whether each token began with either the /g/ or /k/ sound by pressing an appropriately labeled key. Participants were instructed to make their responses as quickly as possible without sacrificing accuracy and to guess if they were unsure. The ISI was 1000 ms, timed from the participant’s response to the onset of the next auditory stimulus. Participants were given a brief break in between the two blocks. The entire procedure took approximately 20 minutes.

2.2 **Results**

Figure 3 shows mean proportion /k/ responses as a function of VOT for listeners in each order group for the same talker and different talker conditions. Robust distributional learning is observed. Consider first performance for the same talker condition. In Block 1, the identification function is shifted towards shorter VOTs for listeners in the Short-Long compared to the Long-Short order group, consistent with listeners hearing shorter VOT distribution in the former
compared to the latter group. In Block 2, however, this pattern is reversed. Listeners in the Long-Short order group show an identification function that is moved towards shorter VOTs compared to listeners in the Short-Long order group. This is again consistent with exposure in each block; in Block 2, listeners in the Long-Short order group heard input distributions containing shorter VOTs than those in the Short-Long order group.

**Figure 3.** Mean proportion /k/ responses in Experiment 1 as a function of VOT in each block for each order separately for the same (top) and different talker (bottom) conditions. Means reflect grand means calculated over by-subject averages; error bars indicate standard error of the mean.

When inspecting performance between the two talker conditions, no difference between talker conditions is readily apparent, suggesting that learning across the test blocks was equivalent regardless of whether Block 2 consisted of exposure from the same or a different
talker compared to Block 1. Moreover, displacement between the two order groups in Block 2 is attenuated relative to the Block 1; a pattern that holds for both the same and different talker conditions.

To examine this pattern statistically, trial-level responses (/g/ = 0, /k/ = 1) were fit to a generalized linear mixed-effects model (GLMM) with the binomial response family using the glmer() function from the lme4 package (Bates et al., 2015) in R. The model was limited to trials containing VOTs that were presented in both sets of input distributions (40,320 trials out of a total of 48,640 trials) because a failure to do so could lead to order effects emerging in each block solely due to the different range of VOTs presented in each block for each order group. The model contained fixed effects of VOT, block, order, condition, and all interactions among these factors. VOT was entered into the model as a continuous variable, scaled/centered around the mean. Block (Block 1 = -1, Block 2 = 1), order (Short-Long = -1, Long-Short = 1), and condition (Same Talker = -1, Different Talker = 1) were sum-coded. The random effects structure consisted of random intercepts by subject and random slopes by subject for VOT, block, and their interaction; this structure represents the maximal random effects structure given the experimental design.

The model results are shown in Table 2. As expected, there was a main effect of VOT (p < 0.001), indicating that /k/ responses increased as did VOT. Critically, there was a significant interaction between block and order (p < 0.001), indicating that the change in /k/ responses across blocks was not equivalent for each order group (p < 0.001). There was also a significant interaction between VOT, block, and order, indicating that the block by order interaction differed across continuum steps (p = 0.003). There was no main effect of condition, nor did condition interact with block and/or order.
Table 2. Results of the generalized linear mixed-effects model for Experiment 1. The model contained 40,320 observations total across 160 participants. All test statistics reflect those reported by the glmer() function.

<table>
<thead>
<tr>
<th>Fixed effect</th>
<th>$\hat{\beta}$</th>
<th>SE</th>
<th>$z$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>1.326</td>
<td>0.095</td>
<td>14.005</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>VOT</td>
<td>2.815</td>
<td>0.084</td>
<td>33.330</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Block</td>
<td>0.094</td>
<td>0.054</td>
<td>1.740</td>
<td>0.082</td>
</tr>
<tr>
<td>Order</td>
<td>-0.139</td>
<td>0.094</td>
<td>-1.475</td>
<td>0.140</td>
</tr>
<tr>
<td>Condition</td>
<td>0.077</td>
<td>0.094</td>
<td>0.817</td>
<td>0.414</td>
</tr>
<tr>
<td>VOT * Block</td>
<td>0.108</td>
<td>0.048</td>
<td>2.247</td>
<td>0.025</td>
</tr>
<tr>
<td>VOT * Order</td>
<td>-0.196</td>
<td>0.083</td>
<td>-2.362</td>
<td>0.018</td>
</tr>
<tr>
<td>Block * Order</td>
<td>0.421</td>
<td>0.053</td>
<td>7.998</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>VOT * Condition</td>
<td>0.032</td>
<td>0.083</td>
<td>0.386</td>
<td>0.700</td>
</tr>
<tr>
<td>Block * Condition</td>
<td>-0.007</td>
<td>0.053</td>
<td>-0.135</td>
<td>0.893</td>
</tr>
<tr>
<td>Order * Condition</td>
<td>-0.052</td>
<td>0.094</td>
<td>-0.553</td>
<td>0.581</td>
</tr>
<tr>
<td>VOT * Block * Order</td>
<td>0.136</td>
<td>0.045</td>
<td>3.000</td>
<td>0.003</td>
</tr>
<tr>
<td>VOT * Block * Condition</td>
<td>-0.029</td>
<td>0.045</td>
<td>-0.650</td>
<td>0.516</td>
</tr>
<tr>
<td>VOT * Order * Condition</td>
<td>-0.045</td>
<td>0.083</td>
<td>-0.547</td>
<td>0.584</td>
</tr>
<tr>
<td>Block * Order * Condition</td>
<td>0.030</td>
<td>0.053</td>
<td>0.574</td>
<td>0.566</td>
</tr>
<tr>
<td>VOT * Block * Order * Condition</td>
<td>-0.017</td>
<td>0.045</td>
<td>-0.383</td>
<td>0.701</td>
</tr>
</tbody>
</table>

In order to explicate the nature of the block by order interaction, the emmeans package was used to test all pairwise comparisons involving block and order, applying a Bonferonni correction to $p$-values in order to control family-wise error rate. With respect to the within-subjects paired comparisons, listeners in the Short-Long group had fewer /k/ responses in Block 1 compared to Block 2 ($\hat{\beta} = 0.654$, $SE = 0.152$, $z = 4.295$, $p < 0.001$) and listeners in the Long-Short group had fewer /k/ responses in Block 2 compared to Block 1 ($\hat{\beta} = -1.030$, $SE = 0.150$, $z = -6.879$, $p < 0.001$). These comparisons confirm that both order groups showed evidence of
distributional learning. With respect to the between-subjects comparisons, there were fewer /k/ responses in the Short-Long compared to the Long-Short order group ($\beta = 1.120, SE = 0.219, z = 5.107, p < 0.001$) in Block 1 and were fewer /k/ responses in the Long-Short compared to the Short-Long order group ($\beta = -0.564, SE = 0.212, z = -2.664, p = 0.046$) in Block 2. However, the order effect in Block 2 is marginally reliable ($p = 0.046$) and shows a smaller effect size (as indexed by the absolute beta estimate) than the order effect in Block 1. This pattern suggests that the interaction observed in the omnibus model reflects an attenuation of the order effect in Block 2 compared to Block 1.

To promote more direct comparison to the predictions generated by the computational simulations, the VOT voicing boundary was calculated for each participant in each block. To do so, trial-level responses (0 = /g/, 1 = /k/) in each block (for each participant) were fit to a logistic regression with VOT as a fixed effect. The parameters of the resulting logistic regression were used to locate the voicing boundary, defined as the VOT corresponding to 0.50 /k/ responses, according to the equation (4).

$$\beta_0 + \beta_1 X = \log\left(\frac{0.5}{1-0.5}\right)$$

Figure 4 shows the distribution of derived boundaries averaged across participants. Consistent with the results of the GLMM and subsequent paired comparisons, the displacement in voicing boundaries between the two order groups is larger in Block 1 compared to Block 2, and similar patterns are observed between the same and different talker conditions.
The results of Experiment 1 provided no evidence to suggest that distributional learning was linked to individual talkers’ voices. Instead, listeners in both the same talker and different talker conditions showed sensitivity to changes in the input statistics to an equivalent degree, with the pattern of learning more consistent with the same talker predictions generated by the computational simulations. Recall that to simulate predictions for the same talker condition, trial-level input was cumulatively integrated with the initial prior specification across both blocks. The simulations predicted an attenuation of the order effect in Block 2 compared to Block 1, reflecting shared exposure between the two order groups over time. The results of Experiment 1 are consistent with this qualitative pattern; however, the two order groups did not completely converge in Block 2 as predicted by the model simulations. Experiment 2 was conducted as a replication of Experiment 1. Methodological procedures followed Experiment 1 exactly except for switching the assignment of talkers TD and TY in the same and different talker conditions.

Figure 4. Boxplots for Experiment 1 showing the distribution of derived category boundaries across participants in each order group for each block, separately for the same talker and different talker conditions.
3.1 Methods

Participants. Participants (n = 160; 84 women, 76 men)² were recruited from the Prolific participant pool (www.prolific.co) following the demographic constraints described for Experiment 1. Participants were randomly assigned to one of four experimental groups formed by crossing two levels of order (Short-Long vs. Long-Short) and two levels of talker variability (same talker vs. different talker). All participants provided informed consent following procedures approved by the University of Connecticut Institutional Review Board.

Stimuli. The stimuli for Experiment 2 were the same as those used for Experiment 1.

Procedure. The exact same procedure as outlined in Experiment 1 was used here except for assignment of the two talkers to the same and different talker conditions. Listeners in the same talker groups heard talker TY in both blocks. Listeners in the different talker groups heard talker TD in the short-VOT input block and talker TY in the long-VOT input block.

3.2 Results

Figure 5 shows mean proportion /k/ responses as a function of VOT for listeners in each order group in each block, separately for the same talker and different talker conditions. As observed in Experiment 1, there is robust evidence of distributional learning. In all four panels, the identification response functions are displaced in line with the distributional input presented in each block/order condition. In addition, the displacement between order groups in Block 2 is attenuated relative to Block 1, and no discernable differences are observed between the same talker and different talker conditions.

To examine this pattern statistically, trial-level responses (/g/ = 0, /k/ = 1) were fit to a

---

² An additional 38 participants were tested but excluded due to failure to pass the headphone screen, failure to exhibit a logistic response function in one or both of the test blocks, and/or showing a category boundary exceeding 40 ms of the intended category boundary in each block.
generalized linear mixed-effects model (GLMM) with the binomial response family using the glmer() function from the lme4 package (Bates et al., 2015) in R following the procedure outlined for Experiment 1. VOT was entered into the model as a continuous variable, scaled and centered around the mean; block, order, and condition were sum-coded. The maximal random effects structure for the design was used, consisting of random intercepts by subject and random slopes by subject for VOT, block, and their interaction.

**Figure 5.** Mean proportion /k/ responses in experiment 2 as a function of VOT in each block for each order separately for the same (top) and different talker (bottom) conditions. Means reflect grand means calculated over by-subject averages; error bars indicate standard error of the mean.

The model results are shown in Table 3. As expected, there was a main effect of VOT (p < 0.001), indicating that /k/ responses increased as did VOT. Critically, there was a significant
interaction between block and order (p < 0.001), indicating that the change in /k/ responses
across blocks was not equivalent for each order group (p < 0.001). As in Experiment 1, there was
no main effect of condition, nor did condition interact with block and/or order.

### Table 3. Results of the generalized linear mixed-effects model for experiment 2. The model contained 40,320 observations total across 160 participants. All test statistics reflect those reported by the glmer() function.

<table>
<thead>
<tr>
<th>Fixed effect</th>
<th>$\hat{\beta}$</th>
<th>SE</th>
<th>z</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>1.313</td>
<td>0.088</td>
<td>14.862</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>VOT</td>
<td>2.872</td>
<td>0.083</td>
<td>34.600</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Block</td>
<td>0.041</td>
<td>0.046</td>
<td>0.890</td>
<td>0.373</td>
</tr>
<tr>
<td>Order</td>
<td>-0.063</td>
<td>0.088</td>
<td>-0.713</td>
<td>0.476</td>
</tr>
<tr>
<td>Condition</td>
<td>-0.101</td>
<td>0.088</td>
<td>-1.155</td>
<td>0.248</td>
</tr>
<tr>
<td>VOT * Block</td>
<td>-0.016</td>
<td>0.050</td>
<td>-0.313</td>
<td>0.754</td>
</tr>
<tr>
<td>VOT * Order</td>
<td>0.003</td>
<td>0.082</td>
<td>0.038</td>
<td>0.970</td>
</tr>
<tr>
<td>Block * Order</td>
<td>0.294</td>
<td>0.044</td>
<td>6.617</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>VOT * Condition</td>
<td>-0.173</td>
<td>0.082</td>
<td>-2.118</td>
<td>0.034</td>
</tr>
<tr>
<td>Block * Condition</td>
<td>-0.042</td>
<td>0.044</td>
<td>-0.950</td>
<td>0.342</td>
</tr>
<tr>
<td>Order * Condition</td>
<td>0.004</td>
<td>0.088</td>
<td>0.040</td>
<td>0.968</td>
</tr>
<tr>
<td>VOT * Block * Order</td>
<td>-0.014</td>
<td>0.048</td>
<td>-0.296</td>
<td>0.768</td>
</tr>
<tr>
<td>VOT * Block * Condition</td>
<td>-0.022</td>
<td>0.048</td>
<td>-0.463</td>
<td>0.643</td>
</tr>
<tr>
<td>VOT * Order * Condition</td>
<td>0.004</td>
<td>0.082</td>
<td>0.052</td>
<td>0.958</td>
</tr>
<tr>
<td>Block * Order * Condition</td>
<td>-0.061</td>
<td>0.044</td>
<td>-1.368</td>
<td>0.171</td>
</tr>
<tr>
<td>VOT * Block * Order * Condition</td>
<td>-0.021</td>
<td>0.048</td>
<td>-0.431</td>
<td>0.667</td>
</tr>
</tbody>
</table>

In order to explicate the nature of the block by order interaction, the emmeans package
was used to test all pairwise comparisons involving block and order, applying a Bonferonni
correction to p-values in order to control family-wise error rate. With respect to the within-
subjects paired comparisons, listeners in the Short-Long group had fewer /k/ responses in Block
1 compared to Block 2 ($\hat{\beta} = 0.506$, SE = 0.127, $z = 3.982$, $p < 0.001$) and listeners in the Long-
Short group had fewer /k/ responses in Block 2 compared to Block 1 ($\beta = -0.669$, $SE = 0.127$, $z = -5.253$, $p < 0.001$). These comparisons confirm that both order groups showed evidence of distributional learning. With respect to the between-subjects comparisons, there were fewer /k/ responses in the Short-Long compared to the Long-Short order group ($\beta = 0.713$, $SE = 0.193$, $z = 3.688$, $p = 0.001$) in Block 1. However, the order effect in Block 2 was not reliable ($\beta = -0.462$, $SE = 0.201$, $z = -2.305$, $p = 0.127$). As in Experiment 1, the order effect in Block 2 showed a smaller effect size (as indexed by the absolute beta estimate) than the order effect in Block 1. This pattern suggests that the interaction observed in the omnibus model reflects an attenuation of the order effect in Block 2 compared to Block 1.

To promote more direction comparison to the predictions generated by the computational simulations, the VOT voicing boundary was calculated for each participant in each block following the procedure outlined for Experiment 1. Figure 6 shows the distribution of derived boundaries averaged across participants. Consistent with the results of the GLMM and subsequent paired comparisons, the displacement in voicing boundaries between the two order groups is larger in Block 1 compared to Block 2, and similar patterns are observed between the same and different talker conditions.
The remarkable ability to rapidly, accurately, and confidently perceive speech sounds given highly variable acoustic input is something that the majority of people take for granted. The mechanisms through which this occurs has been a topic of research for decades. Requiring the coordinated performance of all the bits and pieces of the sensory hardware and the distinctly flexible software of the processing systems, it is a wonder that a highly variable acoustic signal, such as speech, is understood at all. A long-held theory of speech perception is that listeners form generalized representations of speech sounds they hear in their everyday environments. This generalized form can be used to allow the system to rapidly recognize sounds that vary minimally from the abstraction. However, our acoustic environments are rarely stable, and thus we require the system to be able to adapt to, sometimes significant, changes in the speech produced by our interlocutors.

Past research has shown that the mechanisms responsible for processing acoustic input...
are sensitive to the statistical relationships of speech cues, ranging from distributional information of lower-level spectral and temporal structures to transitional probabilities of higher-level structures (Aslin, Saffran, & Newport, 1998; Clayards et al., 2008; Kleinschmidt & Jaeger, 2015; Weatherholtz & Jaeger, 2016; Weiss, Gerfen, & Mitchel, 2009). Of specific importance to this manuscript, the usage of distributional information relative to low-level spectral and temporal acoustic structures have been a topic of interest in recent years. It is speculated that listeners can adapt to this type of information at varying levels of specificity (Kleinschmidt & Jaeger, 2015; Weatherholtz & Jaeger, 2016). How listeners explicitly adapt to this information has been described by a recent computational account of distributional learning (Kleinschmidt & Jaeger, 2015). This account details how distributional information is updated according to both observation and expectation. Of critical importance to this computational account is whether or not distributional information is tracked and updated in a talker-specific manner.

Previous research has provided evidence that listeners appear to track distributional VOT cues in a cumulative fashion within a given talker (Theodore & Monto, 2019). Other research in adaptation to dimension-based statistics concerning $F_0$-VOT cue relationships shows that listeners adapt to localized, short-term, changes in cue-relationships within a specific talker (Idemaru & Holt, 2020). The adaptation to changes in cue relationships reflect adaptation to changes in informativity over time; as one cue of the cue-pair becomes less informative, listeners down-weight their usage of it (Idemaru & Holt, 2011; Idemaru & Holt, 2014; Liu & Holt, 2015; Idemaru & Holt, 2020). Generalization of this learned down-weighting has been observed in a novel talker, but only if there is sufficient spectral overlap between the exposure and novel talkers (Liu & Holt, 2015). This would suggest that in order for talker-specific adaptation to statistical information to occur there would have to be sufficient difference in the structures...
across the talkers in question. The current study examined the degree to which the iterative updating to cumulative input distributions observed in Theodore and Monto (2019) is talker-specific. Two main findings emerged.

First, the results of the current study provide no evidence that distributional learning for VOT input statistics was talker-specific. We observed no evidence of either a main effect or interaction of talker variability in either experiment, suggesting that the observed behavior was the same regardless of whether they heard the same talker or a different talker in the second block of exposure. Second, distributional learning appears to reflect a cumulative integration of input statistics over time, but to a lesser degree than was predicted by our simulations with the Bayesian belief-updating model of speech adaptation. For both the same talker and different talker conditions, there was always a reliable difference of percent /k/ responses between order groups within the first block, but in the second block, there was a measurable attenuation of the difference between order groups. In Experiment 1, the difference, while statistically reliable, was marginally so and had a smaller effect size compared to the first block. In experiment 2, the difference between order groups was not statistically reliable in the second block, maintaining the pattern of marginal reliability and a smaller effect size compared to the first block that was observed in Experiment 1. The attenuation of the between-subjects difference in block two was observed in tandem with a significant change in within-subject behavioral responses from the first exposure block to the second exposure block. Recall, that the cumulative (i.e., same talker) predictions detailed a change of behavioral responses across blocks within only the Short-Long group, while responses were predicted to remain equivalent across blocks within the Long-Short group and the local (i.e., different talker) predictions detailed a change of behavioral responses across blocks within both order groups. With these results in mind, implications of each finding
are discussed in turn.

Contrary to our predictions, we observed no evidence to indicate talker-specific
distributional learning. This is in contrast to talker-specificity observed across many phenomena
of speech perception including lexically-guided perceptual learning (Eisner & McQueen, 2005;
Eisner & McQueen, 2006; Kraljic & Samuel, 2005), statistical structure adaptation (Weiss,
Gerfen, & Mitchel, 2009), accent adaptation (Trude & Brown-Schmidt, 2012), and speeded word
classification and recognition (Goldinger, 1998). Paired with the evidence of talker-specificity in
characteristic VOT productions (Chodroff & Wilson, 2017; Theodore et al., 2009) and that
listeners are able to track and generalize characteristic productions (Allen et al., 2004; Theodore
& Miller, 2010; Theodore, Myers, & Lomibao, 2015) when exposed to a talker’s modal VOTs,
there is a large precedent that listeners would track VOT distributions in a talker-specific
manner. That said, research regarding the degree of informativity that cues contain may provide
some insight as to why observed a lack of talker-specific distributional learning. When you only
take into account explicit VOT values, there is only a minimal advantage to tracking these values
in a talker-specific manner compared to favoring gender-specific or marginalized distributions
(Kleinschmidt, 2019). While the advantage in doing so is present, there is much more of an
advantage in tracking spectral information in a manner more specific to a talker than
marginalized distributions (Kleinschmidt, 2019). The lack of an observed effect of talker-
specificity on distributional learning in our study may be due to the distributions being tracked
over a marginalized or gender-specific cue distribution. This may be a result of listeners being
unaware of whether or not they would hear productions from the same talker in the second block
and both our talkers being female.

Furthermore, the studies that report talker-specificity effects in speech perception
overwhelmingly observe these effects using stimuli that critically differ in spectral acoustic-phonetic cues across talkers of different genders (Eisner & McQueen, 2005; Goldinger, 1998; Trude & Brown-Schmidt, 2012; Weiss, Gerfen, & Mitchel, 2009). Eisner and McQueen (2005) found that the lexically-guided perceptual learning effect did not transfer to ambiguous fricatives produced by a novel talker. Critically though, the novel talker was male, while the trained talker was female. Trude and Brown-Schmidt (2012) found that adaptation to the presence or absence of Chicago accent (/æ/-/ɛ/ before /g/ rather than /æ/-/ɛɪ/) was found to be talker-specific. However, their accented and unaccented talkers were male and female respectively, providing additional cues by which listeners may be driving the observed effect. Weiss, Gerfen, and Mitchel (2009) found that listeners were only able to successfully learn incongruent statistical structures when they were spoken by different talkers. As observed in the previously mentioned studies, the two talkers in Weiss, Gerfen, and Mitchel (2009) differed in gender, one was male and the other was female. As our study stands, we cannot make explicit claims as to whether or not the presented cue distributions were tracked in a gender-specific manner, we can only state that there was a distinct lack of talker-specificity. The theoretical instantiation of the Bayesian belief-updating model (Kleinschmidt & Jaeger, 2015) allows for cue distributions to be tracked independently by varying levels of context (e.g., language, gender, talker). As researchers, we need to be wary in claiming talker-specificity of some perceptual phenomena when the evidence of this is only observed across two talkers of differing gender.

Rather than observing talker-specific distributional learning, we observed evidence of listeners integrating distribution information in a cumulative manner regardless of whether the talker remained constant across the two exposure blocks. That is – for both talker variability conditions – identification functions for the two order groups showed greater displacement in
Block 1 compared to Block 2, and thus followed the qualitative predictions of our single talker simulations. However, the observed behavior does not exactly match the single talker simulations and instead falls somewhere in between that of our single talker and different talker simulations. Namely, listeners’ responses display a significant shift from block one to block two within each order group for both the single talker and different talker conditions, but this pattern of movement for both order groups was only predicted for the different talker (i.e., local) simulations. The same talker (i.e., cumulative) simulations only predicted a change in behavior from the first to second block in the Short-Long order group, not the Long-Short order group. However, as noted previously, despite the change in behavior from block one to block two for all order groups, we also observe a degree of convergence across order groups in Block 2. This convergence is predicted by the same talker (i.e., cumulative) simulations and not the different talker (i.e., local) simulations. Our behavioral results suggest a degree of flexibility in distributional adaptation that is not completely captured by the Bayesian belief-updating model of speech adaptation (Kleinschmidt, 2017) as implemented in the current work. The observation of flexibility to change in cue distributions is not a surprising one however, as it has been observed before by Idemaru and Holt (2011; 2014; 2020). We show similar evidence that listeners shift in response to changes in statistical input, but not to a degree that leads to a complete swap to the newly presented signal statistics (Idemaru & Holt, 2011; Idemaru & Holt, 2014; Idemaru & Holt, 2020). As to why this flexibility was not mirrored in our simulations, consider the following. One of the model parameters, confidence level, determines the degree to which observed information is integrated into the prior distribution. The smaller the confidence level is in the prior, the larger effect integrated observations will have on the posterior. The presented simulations were run at a confidence level equal to 50 for both the /g/ and /k/ prior
distributions. It may be the case that the confidence of the simulations need to be set even lower to get a predicted change over blocks for the Long-Short order group. The pattern reflected in the behavioral responses may also reflect a change in the preferred prior. One of the most crucial choices to make when modeling anything in a Bayesian fashion is to correctly select the prior to be updated. Our prior was selected to be in line with available marginalized VOT production data. However, we cannot accurately predict the prior that listeners enter our experiments with complete certainty. The degree to which listeners vary in how readily they integrate distributional information may not only be influenced by their “readiness” to adapt (i.e., a confidence level), but also by the space that their prior distributions occupy. By presenting distributional input from a female talker, we may be priming listeners to shift into a more gender-specific cue space. The follow-up of more input from another female talker may not sufficiently fall outside the cue space, hence the lack of a measurable change in behavior. Getting an accurate measure of a listeners’ prior category distributions at multiple points is a necessary direction for future research and is critical for understanding the mechanisms behind statistical learning of any form.

In conclusion, the present study provides evidence that listeners display a keen sensitivity to changes in distributional statistics, consistent with a cumulative registration of input over time. However, we found no evidence that distributional learning measured here was specific to a given talker. This extends previous work in distributional learning by providing explicit evidence that distributional learning of VOT information is not always talker-specific. Future research should examine (1) whether distributional learning of non-spectral cues (e.g., VOT) tracks with gender (instead of specific talkers) by examining whether similar patterns would be observed for talkers who differ in their gender, and (2) the degree to which listeners’ prior category distributions influence adaptation, including refinement of ways to estimate category priors.
5 References


6 Appendix

As described in the main text, multiple simulations were performed in order to ensure that our qualitative predictions for the same vs. different talker conditions were robust to variation in parameter specifications. Specifically, nine sets of simulations were performed reflecting three different prior specifications (described in the main text) and three different confidence levels (50, 100, and 200). This appendix consists of three figures that show the resulting simulation output for each prior specific at each confidence level. Figure 7 shows simulation output when the confidence parameter was set to 50, Figure 8 shows simulation
output when the confidence parameter was set to 100, and Figure 9 shows simulation output when the confidence parameter was set to 200. The resulting predictions for same talker vs. different talker conditions were robust across all simulation parameters.

**Figure 7.** Simulation results for confidence = 50 according to three prior specifications (panels A, B, and C). The prior distributions are shown at left in each panel; mean predicted category boundary for the same talker and different talker conditions is shown at right in each panel. Error bars indicate standard error of the mean across the 40 simulated listeners in each group.
Figure 8. Simulation results for confidence = 100 according to three prior specifications (panels A, B, and C). The prior distributions are shown at left in each panel; mean predicted category boundary for the same talker and different talker conditions is shown at right in each panel. Error bars indicate standard error of the mean across the 40 simulated listeners in each group.
Figure 9. Simulation results for confidence = 200 according to three prior specifications (panels A, B, and C). The prior distributions are shown at left in each panel; mean predicted category boundary for the same talker and different talker conditions is shown at right in each panel. Error bars indicate standard error of the mean across the 40 simulated listeners in each group.