HHS Public Access

Author manuscript

Psychol Rev. Author manuscript; available in PMC 2016 February 07.

Published in final edited form as:

Psychol Rev. 2015 April ; 122(2): 148–203. doi:10.1037/a0038695.

Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel

Dave F. Kleinschmidt and

University of Rochester, Department of Brain and Cognitive Sciences

T. Florian Jaeger

University of Rochester, Departments of Brain and Cognitive Sciences and Computer Science

Abstract

Successful speech perception requires that listeners map the acoustic signal to linguistic categories. These mappings are not only probabilistic, but change depending on the situation. For example, one talker's /p/ might be physically indistinguishable from another talker's /b/ (cf. *lack of invariance*). We characterize the computational problem posed by such a subjectively non-stationary world and propose that the speech perception system overcomes this challenge by (1) recognizing previously encountered situations, (2) generalizing to other situations based on previous similar experience, and (3) adapting to novel situations. We formalize this proposal in the *ideal adapter* framework: (1) to (3) can be understood as inference under uncertainty about the appropriate generative model for the current talker, thereby facilitating robust speech perception despite the lack of invariance. We focus on two critical aspects of the ideal adapter. First, in situations that clearly deviate from previous experience, listeners need to adapt. We develop a distributional (belief-updating) learning model of incremental adaptation. The model provides a good fit against known and novel phonetic adaptation data, including perceptual recalibration and selective adaptation. Second, robust speech recognition requires listeners learn to represent the *structured* component of cross-situation variability in the speech signal. We discuss how these two aspects of the ideal adapter provide a unifying explanation for adaptation, talker-specificity, and generalization across talkers and groups of talkers (e.g., accents and dialects). The ideal adapter provides a guiding framework for future investigations into speech perception and adaptation, and more broadly language comprehension.

Keywords

speech perception; generalization; adaptation; statistical learning; hierarchical structure; lack of invariance; non-stationarity

---

In order to understand speech, listeners have to map a continuous, transient signal onto discrete meanings. This process is widely assumed to involve the recognition of discrete linguistic units, such as phonetic categories, words, and sentences. The relative stability with which we usually seem to recognize these units belies the formidable computational

---

Corresponding author: dkleinschmidt@bcs.rochester.edu.

challenge that is posed by even the recognition of the smallest meaning distinguishing sound units (such as a /b/ or /p/). In this paper, we characterize this computational problem and propose how our speech processing system overcomes one of its most challenging aspects, the variability of the signal across different situations (e.g., talkers). This problem is not unique to speech recognition, but is a general property of inferring underlying categories and intentions in a changing (i.e., subjectively non-stationary) world (see references in Qian, Jaeger, & Aslin, 2012). The framework that we propose here thus has broad relevance for understanding how people manage changes in the statistical properties of stimuli across different perceptual and cognitive tasks.

The recognition of phonetic categories is broadly assumed to involve the extraction and combination of acoustic and, if present, visual cues. This is a complex task for several reasons. The speech signal is both transient and typically unfolds at speeds not under the listener's control. Additionally, perceptual cues to phonetic categories are often asynchronously distributed across the speech signal. That means that some cues to a phonetic category contrast are detectable several syllables in advance of the phonetic segment, while at the same time cues following a segment can still be informative (e.g., rhoticity, Heid & Hawkins, 2000; Tunley, 1999). Beyond the extraction of acoustic cues from the speech signal, there are two problems which have puzzled researchers for decades. First, the mapping from cues to phonetic features or phonetic categories is non-deterministic: from the perspective of a listener, phonetic categories form *distributions* over multiple cue dimensions, and these distribution overlap with those of other categories. Notably, even multiple instances of the same phonetic category produced by the same talker in the same phonetic context will have different physical properties (Allen, Miller, & DeSteno, 2003; Newman, Clouse, & Burnham, 2001). One cause for these distributions is noise in the biological systems underlying linguistic production (e.g., motor noise in the articulators). Similarly, the perceptual system itself is noisy: neurons that respond to certain acoustic features do not deterministically fire when that feature is present (Ma, Beck, Latham, & Pouget, 2006). Additionally, the acoustic properties of the environment like background noise can further alter the linguistic signal.

However, arguably the biggest challenge to speech perception is that the mapping from acoustic cues to phonetic categories can vary across situations. A 'situation' could be characterized in terms of an individual talker or a group of talkers with a similar way of speaking, or other aspects of the environment which lead to systematic changes in speaking style (like a noisy bar). For example, different talkers sometimes realize even the same phonetic categories, in the same phonetic context, with dramatically differently cue distributions (e.g., Allen et al., 2003; McMurray & Jongman, 2011; Newman et al., 2001). These differences might arise from fixed, physical differences in, for instance, vocal tract size, but they also arise from variable or stylistic factors like language, dialect, or sociolect (e.g., Babel & Munson, 2014; Johnson, 2006; Labov, 1972; Pierrehumbert, 2003). These differences in the cue-to-category mapping can be substantial. Figure 1 shows the distributions of one of the primary cues distinguishing between /s/ and /ʃ/ as produced by two different talkers. Such between-talker variability means that one talker's "ship" is physically more like another's "sip". This problem is known as the *lack of invariance* and is

one of the oldest problems in speech perception (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967). The focus of this article is how listeners manage to accommodate such systematic variability and achieve robust speech recognition.

Overcoming the lack of invariance: The proposal

In the face of the sort of variability between situations—talkers, in this case—seen in Figure 1, it is natural to wonder how we can understand each other at all. We propose that the answer to this question is three-fold:

1. recognize the familiar,

   generalize to the similar, and

   adapt to the novel

As we discuss below, at least the first and the last of these have been more or less explicitly assumed in previous work, and there is at least preliminary evidence for the second. In a familiar situation, the speech recognition system has a great deal of previous experience to draw on, and by *recognizing* a familiar situation it can take advantage of this previous experience. Recognition of the familiar underlies, for example, talker-specific interpretation of the acoustic signal (Creel, Aslin, & Tanenhaus, 2008; Eisner & McQueen, 2005; Goldinger, 1998; Kraljic & Samuel, 2007; Nygaard & Pisoni, 1998). Similarly, *generalizing* to a novel situation based on similar previous experience means the speech recognition system doesn't have to start from scratch each time a new situation is encountered. For example, such generalization allows us to recognize an accent and adjust our interpretations based on previous experience with similar talkers (Baese-berk, Bradlow, & Wright, 2013; Bradlow & Bent, 2008; Sidaras, Alexander, & Nygaard, 2009). At the same time, novel situations might require adaptation beyond what is expected based on previous experience. For example, when encountering a talker with a novel dialect or accent, the speech recognition system must be prepared to *adapt* rapidly and flexibly.

We propose that all three of these strategies arise from the function that the speech recognition system fulfills (i.e., the typical goals of speech recognition), and that the basic design of this system reflects the fact that it must function efficiently under normal circumstances. Specifically, we propose that the three strategies emerge from the organizational constraints on the speech recognition system imposed by the presence of variability both within a single situation and between situations. These constraints lead naturally to a few conceptual components for the proposed framework. First, because there is variability within a situation, the mapping between cues and categories is inevitably *probabilistic*. This makes speech recognition a problem of inference under uncertainty and implies that a robust speech recognition system must use distributional (statistical) knowledge (Clayards, Tanenhaus, Aslin, & Jacobs, 2008; Feldman, Griffiths, & Morgan, 2009; Norris & McQueen, 2008).

Second, because cue distributions themselves vary—sometimes unpredictably—across situations, the system must be prepared, when necessary, to engage in distributional/statistical learning. This is closely related to the notion of life-long implicit learning

(Botvinick & Plaut, 2004; Elman, 1990; Chang, Dell, & Bock, 2006), as well as statistical learning theories of language acquisition (Feldman, Griffiths, Goldwater, & Morgan, 2013; McMurray, Aslin, & Toscano, 2009; Vallabha, McClelland, Pons, Werker, & Amano, 2007), a connection we return to below.

Third, cue distributions do not vary arbitrarily across situations. Rather, the world is structured. For instance, a listener is likely to encounter a particular familiar talker's cue distributions again, relative to any arbitrary cue distributions, and likewise they are more likely to encounter cue distributions that are similar to those encountered in the past, because of regularities in how talkers vary within a language or more specific grouping like gender, dialect, accent, etc. We propose that in order to take advantage of such structured variability, the speech perception system does not *only* engage in distributional learning. In its most basic sense, this is demonstrated by our ability to recognize previously encountered talkers, and use talker-specific experience to guide speech perception. Going beyond talker-specificity, we will discuss evidence that argues for sensitivity to structure over *groups* of talkers or situations. In a world where speech statistics vary in structured ways, life-long adaptation alone is not sufficient for robust speech perception. A robust speech perception system should take advantage of structure in the world that allows previous experience to inform current processing (for similar reasoning applied to other cognitive domains, cf. Qian et al., 2012). It is, we propose, sensitivity to this structure in the world that underlies recognition of familiar situations and generalization to similar ones.

In this paper, we elaborate on this proposal, review the relevant literature, and develop a framework in the tradition of ideal observer models and normative/Bayesian inference (Anderson, 1990) that, we hope, will help guide future work on speech perception. As we detail below, the proposed framework, which we dub the *ideal adapter*, understands all of (1) to (3) above (i.e., recognition, generalization, and distributional learning) as the result of selecting and adapting the appropriate generative model for the current situation based on the integration of prior and present experience (hence, the name for the proposed framework). This brings a unifying and—at least in parts—formalized computational framework to a set of ideas that have been assumed—more or less explicitly—by others before us. For example, it is widely assumed that speech perception is talker-specific (e.g., Creel & Bregman, 2011; Pardo & Remez, 2006; Pisoni & Levi, 2007) and recent work has begun to investigate our ability to generalize across talkers (e.g., Bradlow & Bent, 2008; Eisner & McQueen, 2005; Kraljic & Samuel, 2007; Sidaras et al., 2009). The ideal adapter framework ties together these different lines of work, emphasizing the crucial roles of both the structure of listeners' prior knowledge *and* their ability to learn the statistics of novel situations. It has so far not been fully recognized, we submit, just how far-reaching the consequences of these two aspects of speech recognition are. Laying out the consequences of these two aspects of the framework thus forms the core of this article.

We begin our exposition in Part I with adaptation to the novel. For this, we focus on situations in which listeners have high certainty (i.e., 'know') that they need to adapt. We formalize the problem of adaptation and test the predictions of the ideal adapter framework through an implemented model. We focus on two well-studied phonetic adaptation phenomena: the first where listeners recalibrate one phonetic category in response to

auditorily ambiguous stimuli *labeled* as one category (perceptual recalibration or phonetic adaptation, Bertelson, Vroomen, & de Gelder, 2003; Kraljic & Samuel, 2005; Norris, McQueen, & Cutler, 2003), and the second where listeners change their classification behavior after repeated exposure to the same prototypical stimulus (selective adaptation, Eimas & Corbit, 1973; Samuel, 1986). This leads us to develop and test novel predictions of the proposed framework. In this part of the paper, we spend a substantial amount of time developing intuitions about the mechanics of how—in the ideal adapter framework—listeners can update their beliefs about the cue distributions in the current situation based on direct experience. In doing so, we illustrate how the proposed perspective relates to and diverges from standard accounts of speech recognition.

In Part II, we turn to situations where previous knowledge is crucial for robust speech perception: recognition of familiar situations and generalization to similar novel situations. In contrast to the flexibility demanded by novel situations, in familiar situations listeners can benefit from *stable* representations of past experience. The ideal adapter framework provides a natural link between the distribution of speech statistics in the world—at the level of individual talkers and groups (e.g., dialect, gender, language)—and different strategies for how listeners can achieve robust speech perception in the face of the lack of invariance. In Part II we will discuss what structure there is in the world that listeners can take advantage of and review the evidence that they *do* take advantage of it. In doing so, we identify directions for future research and isolate a number of specific questions that we consider particularly critical for our understanding of the human speech recognition system.

Finally, we close in Part III by putting the framework we have developed into broader perspective. In particular, we will address how our approach relates to other approaches to the problem of the lack of invariance. Following that, we will discuss how our framework might inform broader issues in speech perception, language comprehension, and more domain-general learning and adaptation. Our approach is a computational-level one and as such compares only indirectly to mechanistic- or algorithmic-level approaches (Marr, 1982), but it nevertheless provides a set of tools for reasoning about speech perception (and language comprehension more generally) which can help sharpen questions for research at other levels. For example, the questions raised by the ideal adapter framework also speak to the debate between episodic, exemplar-based or more abstract phonetic representations (Johnson, 1997a; Goldinger, 1998; McClelland & Elman, 1986; Norris & McQueen, 2008; Pierrehumbert, 2003). They also relate to the acquisition of phonetic categories, which can be seen as another type of distributional learning problem (Maye, Werker, & Gerken, 2002; McMurray et al., 2009; Vallabha et al., 2007), and to language processing at higher levels (e.g., Fine, Jaeger, Farmer, & Qian, 2013; Grodner & Sedivy, 2011; Kamide, 2012; Kurumada, Brown, & Tanenhaus, 2012; Kurumada, Brown, Bibyk, Pontillo, & Tanenhaus, 2014). We also discuss recent research that has found adaptive behavior in language processing above the level of speech perception. At its most basic, the ideal adapter framework also contributes to the burgeoning literature on learning in a variable world (e.g., change detection, Gallistel, Mark, King, & Latham, 2001; hierarchical reinforcement learning, Botvinick, 2012; motor learning, Körding, Tenenbaum, & Shadmehr, 2007). Along with other recent approaches, the ideal adapter stresses that the cross-situational statistics of

the world—though being variable—are *structured*, and that our cognitive systems have evolved to take advantage of this structure.

Part I

The ideal adapter framework

Adaptation in speech perception has received a great deal of attention recently. For example, when listeners initially encounter accented speech, they process it more slowly and less accurately, but this disadvantage dissipates within a matter of minutes (Bradlow & Bent, 2008; Clarke & Garrett, 2004 and references therein). Similarly, listeners rapidly adapt to synthesized and otherwise distorted speech (e.g., Davis, Johnsrude, Hervais-Adelman, Taylor, & McGettigan, 2005). Adaptation is not limited to cases of highly unusual pronunciation, such as foreign accents. Even relatively subtle divergences from standard cue distributions can lead to adaptation. For example, listeners adapt to a talker who produces cue distributions with a typical mean value but less variability than normal (Clayards et al., 2008). This suggests that continuous and implicit adaptation to subtle deviations from auditory expectations is a pivotal component of the human speech perception systems.

What is lacking thus far, however, is a better understanding of *how* and *when* we adapt. Specifically, how do listeners detect that their current linguistic representations are inadequate for the current situation, and how is evidence from the currently processed speech stream integrated with previous experience in order to achieve adaptation? Despite the central importance of the lack of invariance problem to speech perception and language understanding (Liberman et al., 1967; Pardo & Remez, 2006), to date there are few cognitive models of adaptation, and as we discuss below, those that do exist do not link it to other strategies for dealing with the lack of invariance. State of the art models of speech perception have begun to address the non-determinism inherent in the mapping from cues to categories, but ignore or abstract away from the specific contributions of the lack of invariance (Feldman et al., 2009; Feldman, Griffiths, et al., 2013; Norris & McQueen, 2008).[1]

We propose that the first important step is to ask *why* phonetic adaptation occurs at all, or rather why one would expect speech perception to exhibit adaptive properties. To that end it is helpful to understand speech perception as a problem of inference under uncertainty. The acoustic cues that provide information about the talker's intended message are variable and ambiguous, and thus each individual cue is only partially informative. In order to effectively infer the underlying message, information must be integrated from many sources, and as we will discuss below, this must be guided by knowledge of the distribution of cues associated with each linguistic unit. However, because of the lack of invariance, these distributions differ across situations (e.g., talkers).

[1]In work on automatic speech recognition, however, the interest in talker-independent speech recognition has led to a range of proposals that are similar in varying degrees to what we propose here for human speech recognition (e.g., Gauvain & Lee, 1994; Leggetter & Woodland, 1995; Shinoda & Lee, 2001).

Adaptation through belief updating—Our central proposal is twofold. First, listeners do not have direct access to the true distribution but rather uncertain *beliefs* about them based on a limited number of observations. Second, inaccurate beliefs about the underlying distributions can lead to slowed or inaccurate phonetic categorization, and in order to achieve robust speech perception across situations, listeners must adapt. In this view, adaptation reflects a sort of incremental distributional learning, and such distributional learning can be computationally characterized as *belief updating*. This incremental distributional learning has to integrate recent experience with a novel situation with prior knowledge and assumptions about the language. In this sense, the proposed account builds on and expands on the general idea of life-long implicit learning (Botvinick & Plaut, 2004; Chang et al., 2006; Elman, 1990) and that the processing of language input is inevitably tied to implicit learning (e.g., Clark, 2013; Dell & Chang, 2014; Jaeger & Snider, 2013). Unlike much of this work, however, we will argue that the implicit beliefs listeners hold based on previous experience are not unstructured. Rather, they reflect higher-level knowledge (beliefs) about different talkers, groups of talkers, dialect and accents, and so on. We return to this in the second part of the paper.

The first proposition of our framework is that human speech perception relies on a *generative model*, or the listener's knowledge of how linguistic units (words, syllables, biphones, phonetic categories, etc.) are realized by different distributions of acoustic cues. Such knowledge allows for speech perception to proceed by comparing how well each possible explanation—higher-level linguistic unit—*predicts* the currently observed signal. The proposal that language comprehension proceeds via prediction of the signal accounts for a variety of properties of language understanding beyond the ones discussed here (cf. Dell & Chang, 2014; Farmer, Brown, & Tanenhaus, 2013; Jaeger & Snider, 2013; MacDonald, 2013; Pickering & Garrod, 2013) and is closely related to similar proposals from visual perception and other domains (Clark, 2013; Friston, 2005; Hinton, 2007; Y. Huang & Rao, 2011; Rao & Ballard, 1999).

Our second proposition is that the cue values predicted from a given linguistic unit depend not only on what is being said (the phonetic category, biphone, word, etc.) but also on *who* is saying it, and good speech perception depends on using an appropriate generative model for the current talker, register, dialect, etc.. The listener never has access to the *true* generative model, but rather only their uncertain *beliefs* about that generative model. Thus, adaptation can be thought of as an update in the listener's talker- or situation-specific beliefs about the linguistic generative model. The idea that speech adaptation reflects learning about the linguistic generative model is not in and of itself novel, but it has largely been implicit in the empirical literature thus far and our proposal provides an explicit framework and formalization for understanding the link between learning and processing in speech perception.

Our goal is to provide a framework for understanding, on the one hand, how listeners might best represent past experience with different situations, and on the other hand how listeners can integrate that previous experience with evidence from the currently processed speech signal in order to infer an appropriate generative model for each situation. As we discuss below, the listener needs to bring their beliefs about the distribution of cue values for each

category into alignment with the actual distribution that the talker is producing. Because the speech signal unfolds over time, and because the fine-grained acoustic information fades rapidly, this belief updating must be done incrementally. However, this is difficult because each individual speech sound is corrupted by the intrinsic variability of the speech production, transmission, and perception process, and hence not an unambiguous cue to the underlying distribution. That is, when a listener hears a cue value that they do not expect, it could be due either to a change in the underlying distributions, or because deviations from prototypical cue values happen for a variety of other reasons (muscle fatigue, coarticulation, background noise, etc.). The question is thus how the listener should incorporate each new piece of evidence into their beliefs. We address this question by developing an *ideal adapter* framework, which, in the tradition of computational-level/rational analysis (Anderson, 1990; Marr, 1982), sets out the statistically optimal way to do this integration. By this, we mean that adaptation reflects an inference process which combines prior beliefs and recent experience proportional to the degree of confidence in each.[2],

The ideal listener or phonetic categorizer—Our ideal adapter framework builds on a foundation of *ideal listener* models, which describe the problem of speech perception as statistical inference of the talker's intentions (Clayards et al., 2008; Feldman et al., 2009; Norris & McQueen, 2008; Sonderegger & Yu, 2010). Such inference includes inferring intermediate linguistic units, like phonetic categories, either as a means to another end or as an end in itself, as in the case of explicit experimental phonetic categorization tasks. Because we focus specifically on the problem of inferring phonetic categories, for our purposes an ideal listener model is better characterized as an *ideal phonetic categorizer* model. For the broader goal of understanding speech perception, it should, however, be kept in mind that the human listener is not just a phonetic categorization machine, and phonetic categorization typically serves other ends (such as lexical access, Norris & McQueen, 2008, or even the successful inference of communicative intentions, Jaeger & Ferreira, 2013).

Because of the inherent variability of how a phonetic category is realized acoustically, any particular cue value is in principle ambiguous, and thus phonetic categorization is a problem of inference under uncertainty. Such inference can be formally expressed in the language of Bayesian statistics. In the general case, the posterior probability of each category $C = c_i$ after observing cue value x is related to the *prior probability* of c_i, $p(C = c_i)$ and the *likelihood* of x under category c_i, $p(x|C = c_i)$, according to Bayes rule:

$$p(C=c_i|x)=\frac{p(x|C=c_i)p(C=c_i)}{\sum_j p(x|C=c_j)p(C=c_j)}$$

[2]This is not intended as a claim that such inference is resource-free or that there is unlimited memory. Importantly, there are many cognitively plausible algorithms which provably approximate—in principled ways—the type of rational inference assumed here for simplicity's sake, and do so with limited resources (e.g., Sanborn, Griffiths, & Navarro, 2010). Many of these algorithms are closely related to mechanistic models like exemplar models (see Gibson, Rogers & Zhu, 2013; Shi, Griffiths, Feldman, & Sanborn, 2010). We return to this point in the final part of this article.

Because the denominator of the fraction does not depend on the specific category c_i and only serves to ensure that all of the posterior probabilities $p(C = c_i|x)$ sum up to one, it is often omitted and the relationship is written as proportionality:

$$p(C{=}c_i|x) \propto p(x|C{=}c_i)p(C{=}c_i)$$

In this paper, we will begin by addressing a very simplified phonetic categorization problem, in which the listener is trying to decide whether a given cue value is a /b/ or /d/, and later discuss how the approach we develop applies in general. One important cue to this phonetic contrast is the F2 locus, or "target" of the second formant transition (Delattre, Liberman, & Cooper, 1955). Figure 2 shows the spectrograms corresponding to synthesized /aba/ and /ada/ tokens (synthesized by Vroomen, van Linden, Keetels, de Gelder, & Bertelson, 2004). Figure 3 (left panel) shows schematically how the distributions of F2 loci differ for /b/ and /d/: /b/ typically has a lower F2 locus, but there is some variability for both /b/ and /d/. There is thus a continuum from /b/-like to /d/-like F2 locus values.

Let's assume for simplicity's sake that the listener is only considering F2 locus as a cue to the /b/-/d/ contrast. Given an observed F2 locus value, the listener must infer how likely it is that the talker intended to produce the phonetic category C =/b/. That is, what is the posterior probability $p(C = \text{b} \mid \text{F2 locus})$?[3] This quantity is found via Bayes rule:

$$p(\text{b}|\text{F2 locus}) = \frac{p(\text{F2 locus}|\text{b})p(\text{b})}{p(\text{F2 locus}|\text{b})p(\text{b}) + p(\text{F2 locus}|\text{d})p(\text{d})} \quad (1)$$

$$= \frac{p(\text{F2 locus}|\text{b})p(\text{b})}{p(\text{F2 locus})} \quad (2)$$

$$\propto p(\text{F2 locus}|\text{b})p(\text{b}) \quad (3)$$

Bayes rule captures the fact that the posterior probability depends on three things. First, it depends on the *prior probability* of the hypothesis, p(b), which could be higher if /b/ is more frequent in the language than /d/, or if there are other contextually available sources of information that make /b/ more likely, like lexical, visual, or coarticulatory cues. Second, it depends on the *likelihood* p(F2 locus | b), which is the probability of the observed F2 locus value given that /b/ was intended by the talker. Finally, it also depends on how credible *other* hypotheses are, which is really a consequence of requiring that the posterior probabilities of all hypotheses add up to one. This is equivalent to the overall probability of the observed F2 locus value, regardless of which hypothesis is true, and since this quantity is the same for all potential hypotheses it is frequently omitted, as in (3).

For an ideal listener, the probability of recognizing a /b/ should be the estimate of the posterior probability of /b/ (and likewise for /d/) (Clayards et al., 2008; Feldman et al.,

[3]Here and elsewhere, we write "b" for "/b/" in equations, for the sake of brevity. We also write p(b) to indicate $p(C = \text{b})$.

2009). This assumes that the result of speech recognition is not a single category but rather uncertain (or variable) estimates of which categories are more or less likely. This is not a trivial assumption. For example, one might imagine that a listener would improve its categorization accuracy by always 'guessing' the category with the highest probability. However, for speech perception more broadly, there is a benefit to maintaining uncertainty about the category, since additional information often becomes available later in the speech signal (e.g., because of the asynchronous nature of acoustic cues). Indeed, human listeners seem to maintain uncertainty about the speech signal for at least a limited amount of time (cf., right-context effects in word recognition, Bard, Shillcock, & Altmann, 1988; Connine, Blasko, & Hall, 1991; Dahan, 2010; Grosjean, 1985)

Treating speech perception as inference under uncertainty provides substantial insight. Much of this comes from the fact that in such a framework, recognition is accomplished not through purely bottom-up template matching but rather by comparing how well each possible higher-level explanation can *predict* the input signal. This framework provides accounts of effects such as the perceptual magnet effect (Feldman et al., 2009), compensation for coarticulation (Sonderegger & Yu, 2010), and integration of auditory and visual cues (Bejjanki, Clayards, Knill, & Aslin, 2011). It also describes speech and language processing at other levels, including lexical access (Norris & McQueen, 2008), the incremental integration of words into a syntactic parse (Hale, 2001; Levy, 2008a, 2008b), and pragmatic reasoning (M. C. Frank & Goodman, 2012; Goodman & Stuhlmüller, 2013) Moreover, Bayesian inference has been shown to provide a powerful and general computational framework for describing statistically optimal inference under uncertainty, via the integration of prior beliefs and recently observed data. This perspective thus extends to other perceptual and cognitive domains (Griffiths, Chater, Kemp, Perfors, & Tenenbaum, 2010; Kersten, Mamassian, & Yuille, 2004; Tenenbaum & Griffiths, 2001), including sensory adaptation in non-language domains (Fairhall, Lewen, Bialek, & de Ruyter Van Steveninck, 2001; Körding, Tenenbaum, & Shadmehr, 2007; Stocker & Simoncelli, 2006).

One specific advantage of this framework for understanding phonetic adaptation is that it links speech perception behavior with the distribution of cues associated with each category. For an ideal listener, the classification curve for /b/ and /d/ responses is derived directly from the respective posterior probabilities (Figure 3, left panel), which in turn are computed in part from the corresponding likelihood, or distribution, function for each category:

$$p(\text{response}=\text{b}|\text{F2 locus})=p(\text{b}|\text{F2 locus}) \propto p(\text{F2 locus}|\text{b})p(\text{b}) \quad (4)$$

Indeed, listeners do appear to use distributional information in speech perception. Clayards et al. (2008) found that listeners adapt to specific distributions of auditory cues to /b/ and /p/. Listeners in this experiment performed a spoken-word picture identification task, where some of the stimuli were /b/-/p/ minimal pairs like "beach" and "peach". Listeners were randomly assigned to two conditions. In both conditions, the /b/ and /p/ percepts were drawn from normal distributions over the primary acoustic cue to the /b/-/p/ contrast (voice onset timing, VOT, Lisker & Abramson, 1964). In the high-variance condition, the variance around the VOT category means for /b/ and /p/ was large; in the low-variance condition, it was small. Listeners' classification boundaries reflected the distribution of cues that they

experienced: for low-variance exposure, the classification boundaries were steep, while for high-variance exposure the boundaries were shallower, reflecting the greater uncertainty about the intended category that comes with more variable productions of each category. Moreover, the difference in boundary slopes was quantitatively predicted by the difference in the category variances in each case.

This result demonstrates two points. First, by showing that listeners' categorization boundaries reflect the variance of the talker's VOT distributions as predicted by the ideal listener model, they show that listeners are using probabilistic cues in a nearly optimal way. Second, and more importantly for our purposes, they show that listeners are *adapting* to a change in the statistics of these cues. Because they have no experience with the experimental talker before beginning the experiment, any differences in their classification function after exposure reflects something that they have learned about the talker's VOT distributions over the course of the experiment.

The natural question to ask is: *how* do listeners get to the point where distributional information is reflected in their behavior? Intuitively, we might say that coming into a new situation—like an experiment—listeners have some beliefs about the distributions of cues for each phonetic category, and that these beliefs change as the listener gains more experience in that situation. These changes in beliefs about how cues are distributed leads to changes in how any given cue is interpreted, resulting in possibly better comprehension or changes in classification behavior. In the next section we show how—like phonetic categorization—this intuitive idea formally corresponds to statistical inference, but at a different level.

The ideal adapter—Our ideal adapter framework builds on the ideal listener framework described in the last section. The ideal listener depends on distributional information about each category, in the form of the likelihood function $p(x|C)$. We can think of the likelihood function for each category as the listener's prediction about what cue values are likely to occur given that category is produced, and this prediction is used during speech perception to evaluate how well each hypothesized category explains the particular cue value currently being classified. However, we can also think of the likelihood functions as explanations of (and predictions about) the *statistics* of cues for each category. Crucially, these explanations come from the listener's *subjective* knowledge of cue distributions, and likely are not exactly identical to the true statistics of cues in the world, because a listener only has finite observations to work with and thus incomplete information about the true distributions. The consequence of this is that the listener has *uncertain* knowledge of cue distributions.

If the statistics of cues associated with each category were identical or at least similar from one situation to the next, information could be accumulated from *all* observed values to obtain sufficiently accurate—and certain—estimates of the likelihood function. But as discussed above, this is not always the case: talkers can differ dramatically in the acoustic cues they use to realize phonetic categories, and thus the true likelihood function differs across situations (Allen et al., 2003; Labov, 1972; Hillenbrand, Getty, Clark, & Wheeler, 1995; McMurray & Jongman, 2011; Pierrehumbert, 2003).

In order to make good use of bottom-up information from acoustic cues, listeners require the appropriate likelihood function for the current situation. Consider again the case of making a /b/-/d/ decision on the basis of the F2 locus cue, but suppose that we have encountered a new talker—call him Sherman—who produces a different distribution of F2 locus values for /b/, a distribution which is shifted to the right (Figure 3, middle). If the listener continues to use the likelihood function which matches the 'normal' talker Norman's /b/ distribution, comprehension of Sherman's speech will suffer: cue values which were ambiguous for Norman are now more likely to be generated from /b/ (middle bottom; the ideal classification function for Sherman, the solid line, is above the dashed line). Conversely, cue values that are perfectly ambiguous for Sherman (where the solid line crosses $p(\text{b} \mid \text{F2 locus}) = 0.5$) would be much more likely to be /d/ when produced by Norman. That is, a mismatched likelihood function can result in slowed or inaccurate comprehension: inaccurate because the ideal category boundary depends on the likelihood function, and slower because /b/ cue values which are nearly prototypical and highly likely for the new talker would be ambiguous for the standard talker (the resulting uncertainty slows processing in this sort of task; Clayards et al., 2008; McMurray, Tanenhaus, & Aslin, 2002).

Similarly, consider the third talker in Figure 3 (right), Priscilla, whose /b/ productions are substantially more precise than Norman's, resulting in a low-variance cue distribution for /b/, but whose /d/ productions show normal variability. Using Norman's likelihood function to classify Priscilla's productions has similar consequences in this situation: cue values that would have been ambiguous for Norman are now quite a bit more likely to have come from the /d/ distribution because Priscilla's /b/s are so precise.

In both of these situations, comprehension difficulties could be avoided if the listener could use the right likelihood function. If the talker is familiar, this might be as simple as retrieving the right likelihood function based on prior experience with the talker (cf. Goldinger, 1998).[4] But what if the talker has never been encountered before? This is a distributional learning problem: in order to achieve efficient and accurate comprehension of a novel talker, the listener must learn the cue distributions corresponding to the new talker's phonetic categories. This is similar to the problem faced by an infant acquiring their first language, although the adult listener starts with a substantial amount of prior knowledge. Most notably, they know that there *are* different phonetic categories for /b/ and /d/, and that these categories are generally distinguished by the F2 locus cue. As we discuss in the second half of this article, adult listeners also may have experience with similar talkers, providing them with more or less useful previous experience.

Still, inferring the distributions of F2 locus values corresponding to these categories is a difficult task, because the inherent variability in these distributions makes each observed cue value ambiguous as evidence for the underlying distribution: if the observed cue value deviates from the listener's predictions, is it due to inherent within-category variance (which

[4]There is the additional problem of detecting *that* the distributions have changed. Change detection in a probabilistic task is a difficult problem that has so far received little attention in research on speech perception and language processing, but has been investigated for other cognitive domains (for a review of the literature, see Qian et al., 2012) and in research on automatic speech recognition (e.g., Ajmera, McCowan, & Bourlard, 2004; Chen & Gopalakrishnan, 1998). As we outline in the second half of this article, the framework we propose holds the potential of a unified solution to both adaptation and change detection.

will produce *some* outliers), or is it evidence that the predictions themselves—the likelihood functions—are wrong and need to be updated?

Thus, determining the talker's category distributions is a problem of inference under uncertainty, just like the problem inferring the talker's intended category based on an observed cue value, but at another level. That is, in the same way that the listener can use their knowledge of how well each possible category predicts an observed cue value to infer which category is most likely, they can also use knowledge about how well each possible category *distribution* predicts the observed statistics of their recent experience in order to infer which underlying distributions are more or less likely in the current situation. The statistically optimal solution to this inference problem can again be described using Bayes Rule. For simplicity's sake, the cue distribution[5] for a category c can be represented by its mean μ_c and variance σ_c^2. Thus the listener's uncertain *beliefs* about this cue distribution can be represented by a probability distribution over means and variances.

An ideal adapter must infer *both* the category label c *and* and the means and variances of the different underlying categories $\boldsymbol{\mu}, \boldsymbol{\sigma}^2 = \{\mu_c, \sigma_c^2\}$ (where for our example, c = /b/, /d/). Formally, this is expressed by the *joint* posterior distribution over category labels and means and variances, which combines prior beliefs with the likelihood of the observed evidence:[6]

$$p(c, \boldsymbol{\mu}, \boldsymbol{\sigma}^2 | x) \propto \underbrace{p(x|\mu_c, \sigma_c^2)}_{\text{likelihood}} \underbrace{p(c)p(\boldsymbol{\mu}, \boldsymbol{\sigma}^2)}_{\text{prior}} \quad (5)$$

This captures the fact that after observing a cue value x, the listener's joint beliefs about the intended category $C = c$ and the parameters of all categories $\boldsymbol{\mu}$, $\boldsymbol{\sigma}^2$ depend on two things. First, the updated beliefs depend on the likelihood, how well each possible combination of categorization and category parameters can predict the observation x, $p(x|c, \boldsymbol{\mu}, \boldsymbol{\sigma}^2) = p(x|\mu_c, \sigma_c^2)$. Second, they depend on the listener's prior beliefs, both about which categories are most likely to be encountered, $p(c)$, *and* which combinations of category means and variances are most probable, $p(\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$. Both aspects of the prior are based on prior experience. The prior over category probabilities depends on the base rate for each category, as well as its probability in context (based on the surrounding sounds or word, or other acoustic cues besides x),[7] while the prior over category means and variance depends on the sorts of cue distributions the listener has encountered before and expects to encounter again. The nature of the prior over category means and variances is the focus of Part II below. For now all that matters is that the listener thinks some category means and variances are more likely than others.

[5]Here we are representing categories as Gaussian (normal) distributions, because they are both mathematically and intuitively tractable. This assumption is not critical for our purposes. The same logic—of belief updating as inferring distributional properties—applies for any parametric or even non-parametric way of representing the distributions.

[6]Using the shorthand notation of $p(c)$ to indicate the probability that the random variable C has value c, or $p(C = c)$.

[7]The usefulness of context in providing prior information about phonetic categorization is not limited to situations where the categorization of nearby sounds is known with certainty. Simply knowing that sequences of categories which correspond to actual words are more likely than arbitrary strings provides prior information about how each sound is categorized through the joint distribution of categorizations $p(\ldots, c_{i-2}, c_{i-1}, c_i, c_{i+1}, \ldots)$.

The specific way that an ideal adapter updates their beliefs after observing cue value x depends on how they categorize it, and this is captured by the joint posterior distribution $p(c, \boldsymbol{\mu}, \boldsymbol{\sigma}^2 \mid x)$. In general, an observation from category c provides the most evidence about the underlying mean and variance of that category. In the case where the prior beliefs about the parameters of each category are independent of each other, $p(\boldsymbol{\mu}, \boldsymbol{\sigma}^2) = \Pi_c p(\mu_c, \sigma_c^2)$, the beliefs about category c_i are *only* updated if the observation is classified as $C = c_i$.

$$p(\mu_{c_i}, \sigma^2_{c_i} | x, C) \propto \begin{cases} p(x|\mu_{c_i}, \sigma^2_{c_i}) p(\mu_{c_i}, \sigma^2_{c_i}) & \text{if } C = c_i \\ p(\mu_{c_i}, \sigma^2_{c_i}) & \text{if } C \neq c_i \end{cases} \quad (6)$$

In cases where there is uncertainty about how the observation x should be categorized, an ideal adapter should update the beliefs about category c_i as a mixture of the updated beliefs under each possible categorization, weighted by how likely that categorization is overall (averaging or marginalizing over current category parameter beliefs):

$$p(\mu_{c_i}, \sigma^2_{c_i} | x) = \sum_c p(\mu_{c_i}, \sigma^2_{c_i} | x, C = c) p(C = c | x) \quad (7)$$

Again, if we assume that the beliefs about different categories are independent, this mixture consists of two components: one where x is categorized as $C = c_i$ and beliefs about category c_i are updated, and one where it is not and no belief updating occurs:[8]

$$p(\mu_{c_i}, \sigma^2_{c_i} | x) = p(\mu_{c_i}, \sigma^2_{c_i} | x, C = c_i) p(C = c_i | x) + p(\mu_{c_i}, \sigma^2_{c_i}) p(C \neq c_i | x)$$

To return to the example above of classifying a token as either /b/ or /d/ based on F2 locus, the posterior distribution over the mean and variance of /b/ after observing a particular F2 locus value is thus

$$p(\mu_{\mathrm{b}}, \sigma^2_{\mathrm{b}} | \text{F2 locus}) = p(\mu_{\mathrm{b}}, \sigma^2_{\mathrm{b}} | \text{F2 locus}, \mathrm{b}) p(\mathrm{b} | \text{F2 locus}) + p(\mu_{\mathrm{b}}, \sigma^2_{\mathrm{b}}) p(\mathrm{d} | \text{F2 locus})$$

In conversational speech, acoustic observations are often labeled with high certainty, and so $p(C = c_i \mid x) \approx 1$ for some category c_i. Such label information can come both from top-down linguistic context (like phonotactics or lexical disambiguation), or from other bottom-up cues. For example, when distinguishing /b/ from /d/, the closure of the lips during /b/ provides a very informative visual cue, effectively labeling the auditory percept (Vroomen et al., 2004).

In such cases where there is some other source of information that labels the observed F2 locus value as a /b/, the resulting conditional posterior distribution over the mean and variance of /b/ simplifies to:

[8] In the general case where the prior beliefs about different categories' parameters are not independent, the posterior is still a mixture, but is a mixture of beliefs, updated in different ways and to different extents, rather than just updated and non-updated beliefs.

$$p(\mu_b, \sigma_b^2 | \text{F2 locus}, \text{b}) \propto \underbrace{p(\text{F2 locus} | \text{b}, \mu_b, \sigma_b^2)}_{\text{likelihood}} \underbrace{p(\mu_b, \sigma_b^2)}_{\text{prior beliefs}} \quad (8)$$

Here, the relevant prior distribution is just the listener's prior beliefs about the mean and variance of the F2 locus cue for the /b/ category. Likewise, the likelihood considers only how well each combination of /b/ mean and variance account for the observed cue value. Below, we model incremental adaptation for cases where the category labels are known with high certainty (and thus (8) holds). We also assume that the prior beliefs about /b/ and /d/ are independent. We make these assumptions for the sake of simplicity and tractability in modeling, and it is important to keep in mind that they do not represent assumptions of the *framework*, which makes qualitatively the same predictions whether or not these assumptions turn out to be true.

In sum, the ideal adapter framework predicts that optimal phonetic adaptation depends on three things: the statistics of the observed percepts (e.g. their mean and variance), the listener's prior beliefs about the statistics of the relevant categories, and the listener's belief that there is a need to adapt (including their beliefs about the amount of variation in the relevant category across talkers and situations). In the next five sections, we illustrate the role of the first two factors in phonetic adaptation experiments where the third factor is a given (i.e., where there is a clear need to adapt and for which previous work has shown that listeners indeed adapt, Bertelson et al., 2003; Kraljic & Samuel, 2005; Norris et al., 2003). In order to do this we specify a basic belief updating model in the ideal adapter framework which quantifies how the exposure statistics and the listener's prior beliefs about those statistics interact.

With this model in hand, we do five things. First, as a basic evaluation we address the phenomenon of phonetic recalibration or perceptual learning. Such perceptual learning is typically thought to be due to changes in the underlying representations of the adapted categories which generally serves the purpose of robust speech perception, and is naturally accounted for by the ideal adapter framework. In particular, we show that the incremental build-up of recalibration is accounted for by our basic belief updating model.

Second, we illustrate how the way in which the model accounts for the build-up of recalibration potentially sheds light on the underlying processes. Specifically, the model captures the fact that recalibration is often ambiguous between a change in the underlying mean of the category versus a relaxation of the criterion for what counts as an acceptable exemplar, or a change in the variance of the category.

Third, we examine the predictions of this framework for the selective adaptation paradigm, a paradigm which is typically *not* considered to be due to the perceptual learning which serves robust speech perception. However, we show that the belief updating model accounts for the incremental build-up of this phenomenon as well, using very similar parameters as for recalibration.

Fourth, we explore a little studied property of phonetic recalibration that, *prima facie*, would seem to stand in conflict with our hypothesis that adaptation serves robust speech

perception: prolonged, repeated exposure to the exact same stimulus can eventually undo the recalibration effect (Vroomen, van Linden, de Gelder, & Bertelson, 2007). However, we show that not only is this predicted by the ideal adapter framework under a range of conditions, the belief updating model which accounts for the build-up of selective adaptation and recalibration also accounts for the effect of prolonged repeated exposure in each, and does so simultaneously with a single set of parameters.

Fifth, motivated by the potential link between selective adaptation and recalibration suggested by the proposed framework, we present novel data from a web-based perception experiment which tests the predictive power of our model. Specifically, we test adaptation conditions that are intermediate between recalibration and selective adaptation, for which the model predicts a continuum between classic recalibration and selective adaptation responses.

Basic evaluation of the ideal adapter framework: phonetic recalibration

We begin with an illustration of the basic mechanics of the ideal adapter framework. For this we focus on experiments in which listeners are exposed to a novel talker with an unusual realization of a phonetic contrast. There is a great deal of evidence about the *results* of incremental adaptation. Much of it comes from studies of “phonetic recalibration”, or “perceptual learning” (Bertelson et al., 2003; Kraljic & Samuel, 2005; Norris et al., 2003). These studies use a continuum between two sounds, generally constructed by interpolating between prototypical endpoint tokens (e.g. Kraljic & Samuel, 2005; Norris et al., 2003) or by parametrically manipulating a critical acoustic cue which distinguishes between the two categories. For instance, a /b/-/d/ continuum might be constructed by manipulating F2 locus, as described above (Bertelson et al., 2003; Vroomen et al., 2004). During an exposure phase, listeners hear the item from this continuum which is most ambiguous between /b/ and /d/. This auditorily ambiguous segment is paired with information which consistently “labels” or disambiguates it as a /b/. This labeling is achieved via, for example, lexical disambiguation (e.g. replacing the /b/ in *club* with the ambiguous segment, Kraljic & Samuel, 2005; Norris et al., 2003 etc.) or visual disambiguation (pairing the auditorily ambiguous sound with a video of a person articulating a /b/, which results in a visible labial closure unlike articulation of /d/). After exposure, changes to the listener’s classification function are assessed, for example, by means of a classification test over unlabeled sounds drawn from the continuum (e.g. classifying a continuum of sounds from a prototypical /aba/ to a prototypical /ada/ without either lexical or visual disambiguation).

In what follows, we use the notation $x_{c_1c_2}$ to refer to a sound that is auditorily ambiguous between categories c_1 and c_2, and use superscripts to refer to labeled sounds. So, for example, $x^{\mathrm{b}}_{\mathrm{bd}}$ is a sound auditorily ambiguous between /b/ and /d/ which is labeled (disambiguated) as /b/.

Perceptual recalibration rapidly results in shifted category boundaries. For example, after as few as 10 exposures to $x^{\mathrm{b}}_{\mathrm{bd}}$, the /b/ category has ‘grown’: more of the continuum is now classified as /b/, when tested without the labeling information (Vroomen et al., 2007). The opposite shift is observed for exposure to $x^{\mathrm{b}}_{\mathrm{bd}}$. This is illustrated schematically in Figure 4

(A). Similarly rapid perceptual recalibration has been observed along a variety of phonetic contrasts, including vowels (Maye, Aslin, & Tanenhaus, 2008), fricative place of articulation and manner (Kraljic & Samuel, 2005; Norris et al., 2003), and stop consonant place (Bertelson et al., 2003) and voicing (Kraljic & Samuel, 2006). Perceptual recalibration is typically investigated under the assumption that it reflects the same processes that support general accent adaptation. This assumption is not trivial but there is some support that perceptual recalibration is not simply an artifact of the stimuli being presented in isolation (Eisner & McQueen, 2006) or there only being one unusual pronunciation (Reinisch & Holt, 2014). We return to these issues in the next section.

Qualitatively, perceptual recalibration exhibits several properties that are expected under the ideal adapter framework. First, recalibration seems to reflect implicit learning over phonetic contrasts, rather than strategic effects such as response bias (Clarke-Davidson, Luce, & Sawusch, 2008), or weakening of the criterion for what counts as an acceptable example of a category (Maye et al., 2008; but see next section). Recalibration also appears to affect speech perception through changes in sublexical phonetic category representations since perceptual recalibration effects generalize to novel words by the same talker containing the recalibrated segment (McQueen, Cutler, & Norris, 2006).

Second, perceptual recalibration seems to last: when listeners classify tokens from the same talker 12 hours after initial testing, the magnitude of adaptation is the same as right after initial exposure (Eisner & McQueen, 2006). As we discuss in more detail in the second part of this article, longevity of changes in category boundaries (for a particular situation) is expected under our proposal that adaptation serves to make speech perception robust to changes in situation.

More specifically, the qualitative changes in classification boundaries observed during perceptual recalibration is naturally predicted by the ideal adapter framework (Figure 4, B). Take, for example, the case where the listener is exposed to $x_{\mathrm{bd}}^{\mathrm{b}}$ (Figure 4, left). As the listener updates their beliefs about the shifted distribution of cues for /b/, shifting the mean towards the observed cue values, stimuli in the middle of the /b/-/d/ continuum which previously had roughly equal likelihood under either category (and thus are sometimes perceived as /b/ and sometimes as /d/) are now more likely to have resulted from /b/, resulting in more /b/ responses to unlabeled test stimuli, especially in the previously-ambiguous region of the continuum.

Incremental recalibration—It is encouraging that the ideal adapter framework provides a qualitative account of the results of recalibration. But can this framework account for *incremental* changes in behavior? Belief updating is an incremental process, where the listener accumulates information about the talker's cue distributions one observation at a time. The ideal adapter framework thus not only predicts asymptotic classification behavior —after the listener has fully adapted to the talker's cue distributions—but also how their classification behavior changes with each additional piece of evidence. Unfortunately, few studies have investigated the *incremental* effects of exposure to a novel distribution of sounds, such as would be typical for a new talkers.

A notable exception is Vroomen et al. (2007). Listeners in their study were exposed to repetitions of an audio-visual adaptor, which was composed of a video recording of a talker articulating either /aba/ or /ada/, dubbed with an audio item from a 9-item, synthetic /aba/ ($x_b = 1$) to /ada/($x_d = 9$) continuum. The audio component for each participant was the continuum item that was most ambiguous, x_{bd}. The most ambiguous item was determined during a pre-test block of 98 trials where the entire /aba/-/ada/ continuum was classified.[9]

Instead of the typical recalibration procedure, where exposure and test are separated, Vroomen et al. (2007) measured the degree and direction of adaptation by interspersing audio-only test blocks throughout each exposure block, after 1, 2, 4, 8, 16, 32, 64, 128, and 256 cumulative exposures to the audio-visual adaptor. Specifically, they measured the average proportion of /b/ responses to six-trial test blocks (the three most ambiguous items $\{x_{bd} - 1, x_{bd}, x_{bd} + 1\}$ each repeated twice).

Each participant completed sixteen exposure blocks of 256 exposures. Half of the exposure blocks used a /b/ audio-visual stimulus for exposure, and the other half used a /d/. Of the /b/-exposure blocks, half of these used the auditorily ambiguous stimulus as described above, while the other half used the prototypical /b/ endpoint of the acoustic continuum ($x = 1$), and likewise for the /d/-exposure blocks.[10] Because our goal is to illustrate the workings of the ideal adapter framework when the listener has little prior experience that might be relevant for the current situation, we focus on the first 64 exposures of the first block from each participant (we return to the issue of extended exposure below).

Figure 5 shows the results of Vroomen et al. (2007) demonstrating the build-up of recalibration in the first 64 critical exposures in the first exposure block. The top panel shows the proportion /b/ responses for x_{bd}^{b} and x_{bd}^{d} adaptors separately; more /b/ responses after /b/ exposure (solid line) indicates recalibration, and vice-versa for /d/ exposure (dashed line). A natural measure of the degree of recalibration is thus the *aftereffect* difference score between /b/ and /d/ exposure. A positive aftereffect indicates that /b/ exposure increased /b/ responses, and /d/ exposure *decreased* /b/ responses (and increased /d/ responses), which corresponds to recalibration (Figure 5, bottom).

Recalibration builds up rapidly but incrementally over the first 64 exposures. As discussed above, this build up follows naturally from the ideal adapter framework, as each exposure to the auditorily ambiguous adaptor contributes to a shift in the listener's estimate of the category mean. Next, we quantify and test this prediction.

Modeling build-up of recalibration—In order to evaluate the ability of the ideal adapter framework to account for the results of Vroomen et al. (2007) on the incremental build-up of recalibration, we implemented a basic Bayesian belief updating model based on the principles of the ideal adapter framework.

[9]There were 14 repetitions of each stimulus, except for the two endpoints $x = 1$ and $x = 9$ which were repeated only 6 times and the next most prototypical items $x = 2$ and $x = 8$ which were repeated 8 times each

[10]The prototypical exposure blocks are discussed later.

Bayesian belief updating model: We used a mixture of Gaussians as the underlying model of phonetic categories, where each phonetic category $c \in \{b, d\}$ corresponds to a normal distribution over cue values x with mean μ_c and variance σ_c^2 (e.g. Figure 3, top). Thus, the likelihood of observation x under category c is

$$p(x_i|c, \boldsymbol{\mu}, \boldsymbol{\sigma}^2) = \text{Normal}(\mu_c, \sigma_c^2) \quad (9)$$

The listener's uncertain beliefs about phonetic categories are captured by additionally assigning probability distributions to the means μ_c and variances σ_c^2 of each phonetic category. The prior distribution $p(\mu_c, \sigma_c^2)$ represents the listener's beliefs about category c before exposure to the experimental stimuli, and the posterior $p(\mu_c, \sigma_c^2|X)$ captures the listener's beliefs after exposure to stimuli $X = (x_1, x_2, \ldots, x_N)$ which are known to come from category c(which means that the category labels $C = c$ are known). These two distributions are related via Bayes' Rule:

$$p(\mu_c, \sigma_c^2|X, C{=}c) \propto p(X|\mu_c, \sigma_c^2, C{=}c)p(\mu_c, \sigma_c^2) \quad (10)$$

We used a conjugate prior for the Normal distribution with unknown mean and variance (Appendix A), and this prior distribution has two types of hyperparameters.[11] The first set of hyperparameters captures the prior expected values of the means and variances. The second set of hyperparameters captures the degree of confidence, or, conversely, uncertainty associated with the prior beliefs. Put differently, they determine how much current observations are weighted against previous experience in determining beliefs about the category distributions. In this model, there are two different degrees of confidence: one for the category means, denoted κ_0, and the other for the category variances, denoted ν_0 (see the Appendix for details). An intuitive interpretation of these hyperparameters is as the effective sample size of the prior beliefs. For instance, if the category mean confidence parameter is $\kappa_0 = 10$, then after ten *new* observations the listener's beliefs about the category mean to equally reflect previous and current experience. With fewer than ten new observations, the listener's beliefs about the category mean will be dominated by the mean expected based on previous experience; with more than ten new observations the beliefs about the mean will be increasingly dominated by the mean of the new observations. These hyperparameters thus capture the gradient trade-off between prior experience and current experience in determining the listener's beliefs about phonetic categories. They can be thought of as *pseudocounts* or the number of prior experiences that are relevant for the current situation.[12]

Finally, it is not a priori obvious whether adaptation occurs at the level of auditory cues individually or at some higher level where information is integrated from multiple auditory and/or visual cues. Thus the model includes a third hyperparameter, w which determines the

[11] We use the term *hyperparameters* for terminological clarity to distinguish the *model* parameters from the *category* distribution parameters—means and variances—whose prior and posterior distributions are defined by the model parameters, the hyperparameters.

[12] Even though comparable fits can be obtained using a single, overall effective prior sample sizes (e.g. $\kappa_0 = \nu_0$), the two were fit as separate in order to evaluate the extent to which recalibration was primarily driven by a shift in the category mean or a change in variance.

weight given to the visual cue value in determining the percept. This hyperparameter ranges from $w = 0$ (perceived cue value is not influenced by the visual cue) and $w = 1$ (perceived cue value entirely determined by the visual cue). Adaptation over integrated cues might arise because the listener attempts to infer the talker's *intended* production based on multiple partially informative cues (including top-down category distribution information, Feldman et al., 2009). For more discussion, refer to the Appendix.

Model fitting: The hyperparameters were fit in a two-step process, which is described in detail in the Appendix. The first step is to fix the expected prior means and variances based on the classification curves measured during pre-test, before exposure to the audiovisual adaptor. These hyperparameters are thus not free parameters of the model, in that they are not adjusted to improve the fit to the actual adaptation test data.

The second step is to estimate the three free hyperparameters (i.e., the effective prior sample sizes ν_0 and κ_0 and the visual cue weight w) based on the actual adaptation data. The posterior distribution of the free hyperparameters was obtained using MCMC sampling with a weakly informative prior (to ensure a proper posterior, Gelman, Carlin, Stern, & Rubin, 2003). For further details, we refer to Appendix A. Because of the limited amount of data from each participant (only six trials per test block), we chose to fit the model to the aggregate data from all participants (see Appendix A for motivation).

Results and discussion—The model's fit against the data is shown in Figure 6. The predicted responses are plotted on the aftereffect scale for better comparison to Vroomen et al. (2007). The model effectively captures the qualitative fact that recalibration leads to positive aftereffects which build up incrementally, and provides a quantitatively good fit as well ($r^2 = 0.96$). Specifically, the model captures the fact that recalibration starts off relatively weak and gradually becomes stronger, before eventually leveling off. Thus, not only is the qualitative result of recalibration—a positive aftereffect—predicted by the ideal adapter framework, the effect of cumulative exposure on the incremental build-up of the effect is also predicted well. This suggests that listeners incrementally integrate each observed cue value with their prior beliefs in a way that is predicted by the ideal adapter framework.

We can draw a number of conclusions from the values of the hyperparameters themselves. The best-fitting hyperparameter values are $\nu_0 = 71$, $\kappa_0 = 11$, and $w = 0.53$. First, relative to the real overall sample size—the number of /b/s and /d/s encountered in the world by a typical English-speaking adult—the best-fitting effective prior sample sizes are *extremely* low. That is, listeners appear to put very little weight on their prior beliefs, adapting very quickly to the shifted cue distribution that they observe while taking slightly longer to adapt to the tight clustering (low variance). This may be surprising at first glance, but it is actually qualitatively *predicted* by the ideal adapter framework. In the ideal adapter framework, whether or not (and how much) a listener adapts depends on how relevant they think their previous experiences are for the current situation. Thus in situations like a recalibration experiment where listeners encounter odd-sounding, often synthesized speech in a laboratory setting, they may have little confidence, a priori, that any of their previous experiences are directly informative. We discuss this point in length in the second half of

this paper, where we elaborate on the crucial role of prior experiences for robust speech perception.

Second, the best-fitting value of the visual cue weight hyperparameter *w* places approximately equal weight on the audio and visual cue values. This means that, according to the best-fitting model, listeners perceive the cue value as not fully ambiguous. This makes an interesting prediction about the effects of extended exposure to the same stimulus that we return to below.

Third, the joint distribution—rather then just the point estimates—of the prior effective sample size hyperparameters (Figure 7) reveal that as long as *one* of these hyperparameters is low—on the same order of magnitude of the number of exposures to the adaptor stimulus—the other confidence hyperparameter can become extremely large and not change the model's predictions so much that the likelihood suffers. This is because there are two ways that the positive aftereffect observed here (and in other recalibration studies) might come about after exposure to an ambiguous adaptor stimulus. This is discussed in more detail next.

Recalibration by category shift or expansion?

One of the advantages of model-fitting using Bayesian methods is that it allows us to evaluate the *range* of model hyperparameters which provide a good fit to the data. For the build-up of recalibration modeled in the previous section, the full posterior distribution over model hyperparameters (effective prior sample sizes and visual cue weight) provides interesting insight into how a human learner might adapt. To illustrate this, we examine the posterior distribution of the prior effective sample sizes for the category means and variances given the behavioral data.

The joint distribution of the two confidence hyperparameters—the mean confidence κ_0 and the variance confidence ν_0—is shown in Figure 7. This distribution covers an extremely wide range of both hyperparameters, although this entire range still results in qualitatively consistent *predictions* about the aftereffect at each level of exposure (Figure 6). The wide range covered by the posterior distribution of hyperparameters is due to the limited amount of data available to the model in this particular case. Note that this is not a problem. It merely reflects that these data do not uniquely constrain the model. A human learner exposed to the same data would face the same problem. Indeed, we will see below that when the model is constrained by further data, the posterior distribution of the hyperparameters will become more narrow.

Of interest is that the posterior is bimodal: there are two ways that belief updating can account for the build-up of recalibration. The best-fitting (MAP-estimate) hyperparameters correspond to a shift in the mean of the adapted category: the prior effective sample size of the mean, κ_0, is less than that of the variance, ν_0, and as a result the prior beliefs about the mean are overcome more quickly than the variance. However, this pattern is only true for roughly half the samples from the joint posterior of the hyperparameters ($p_{\text{MCMC}}(\nu_0 > \kappa_0) = 0.55$). In the other half of the samples, the prior effective sample size for the mean is on

average very high, meaning that the positive aftereffect observed in the data is modeled as a change in the *variance* of the adapted category, with a mean that is essentially fixed.

This combination of hyperparameters—flexible variance and fixed mean—can lead to a positive aftereffect after exposure to auditorily ambiguous but labeled tokens in the following way. If the listener has very high confidence in the mean of each category coming into a new situation, then repeated exposure to an ambiguous segment which is labeled as belonging to one category is best explained by the hypothesis that the talker is producing that category with a high degree of variability. Increasing the variance of the recalibrated category in this way means that more likelihood is assigned by that category to the previously-ambiguous part of the continuum to the recalibrated category, and thus leads to a positive aftereffect.

Thus, a positive aftereffect is qualitatively consistent with both a shift in mean *and* an increase in category variance. Moreover, the joint distribution of hyperparameters fit to the build-up of recalibration observed by Vroomen et al. (2007) show that in the ideal adapter framework, the quantitative effect of cumulative exposure on the build-up of recalibration is also ambiguous in the same way.

Maye et al. (2008) behaviorally investigated a similar question. Specifically, they wondered whether positive aftereffects typically observed in recalibration experiments were really due to a shift in the underlying category, or just a relaxation of the criterion for what counts as a good exemplar of the adapted category. They exposed listeners to vowels that were shifted in a particular direction (e.g. shifting the high vowel /ɪ/ in *wicked* down to the mid vowel /ɛ/ to make *'wecked'*). In a lexical decision task after exposure to such downward shifts, listeners accepted more nonwords that were downward-shifted versions of real words, but not nonwords that were upward-shifted words. This corresponds, in the ideal adapter framework, to a shift in the means of the adapted categories, without a substantial change in the variance.

It may be tempting to conclude based on these results that *all* recalibration effects result from shifting category means. However, the ideal adapter framework predicts that positive aftereffects due to changes in *either* means and variances are possible, in different situations, especially depending on whether the listener has greater confidence in their prior beliefs about category variances or means. This is, to the best of our knowledge, a novel prediction. Since one of our goals is to provide a guiding framework for future work on speech processing and adaptation, we elaborate on this prediction.

When might the listener have greater confidence in the mean of a category rather than its variance, and vice versa? In the ideal adapter framework, the listener's prior beliefs about a category parameter constitute a prediction about the distribution of values they might expect that parameter to take on in the future. The level of confidence in prior beliefs is closely related to the level of variability of a particular category parameter that the listener expects across situations. For cues whose typical values vary across situations (e.g. formant frequencies), an ideal adapter should expect substantial variability in the underlying means,

in order to be prepared to shift their beliefs about category means on that cue when appropriate.

The variance of a particular cue for a particular category is closely related to how reliable that cue is at distinguishing one category from another (Allen & Miller, 2004; Clayards et al., 2008; Newman et al., 2001; Toscano & McMurray, 2010): for two categories with fixed means, increasing the variance of both categories means that their distributions will overlap more and, on average, observing that particular cue will be less informative about the intended category. Thus for a cue which varies in reliability from one situation to the next (with relatively stable category means), the ideal adapter should in general be more likely to adjust category variance than means.[13]

Thus, the ideal adapter framework predicts that there are range of strategies available to the listener for adapting to new talkers. In real-life accent adaptation, there are usually *many* categories and cue dimensions where an accent is unusual, and in some cases these differences can be due to both changes in the cue values typically used to realize a category (the mean) and changes in how reliable a given cue is at distinguishing a category (the variance). In real speech there are many partially informative cues to any given category, and it may be a completely reasonable strategy for the listener to simply decide that a particular cue is uninformative and ignore it (or at least downweight it).

This points to a critical empirical gap in our understanding of speech perception. Most existing work on phonetic adaptation follows one of two approaches. The first approach emphasizes relatively natural conditions and accent variability, where language occurs in context (e.g. sentences) and listeners must adapt to accents that vary along many categories and cue dimensions (Baese-berk et al., 2013; Bradlow & Bent, 2008; Clarke & Garrett, 2004; Sidaras et al., 2009). The second approach is that of perceptual recalibration/learning, which typically presents speech as isolated words or syllables and emphasizes acoustic manipulation of a single category or auditory cue. While perceptual recalibration has been observed when these unusual pronunciations were presented as part of running speech in a story (Eisner & McQueen, 2006) or as words spoken by a talker who has a real foreign accent (Reinisch & Holt, 2014), it is an open question under what conditions listeners downweight cues and when they track changes in mean cue values during naturalistic accent adaptation.

In order to address listeners' ability to adapt to novel talkers based on the statistical properties of their speech as predicted by the ideal adapter framework, we see two potentially fruitful directions. First, we think that perceptual recalibration paradigms should be scaled up to explore the role of natural levels of within-category, within-talker variability and the role of controlled deviation in multiple categories and cue dimensions in recalibration. Second, we think that naturalistic accent adaptation paradigms might be refined to specifically investigate how accent difficulty is driven by deviations—from unaccented speech—in the *average* value of a cue versus unusual *variability* in that cue.

[13]This is not to say that an ideal adapter would *not* adapt to changes in category distributions for cues whose means or variances do *not* vary much across situations. Rather, the amount of variability in a particular category's statistics over situations combines with the listener's overall level of confidence that their prior experience is relevant for the current situation.

Given that, across talkers, the average value of some cues varies quite a bit (Newman et al., 2001), while for others it is relatively consistent (Allen et al., 2003), it might be expected that listeners will have a harder time adapting to accented speech which is characterized by deviant values of cues that are typically stable across talkers (like VOT Sumner, 2011).

We have discussed how phonetic recalibration is qualitatively predicted by the ideal adapter framework, and presented a model in this framework which captures the incremental build-up of recalibration quantitatively. In the next three sections we show how this framework provides a potentially unifying perspective on phonetic adaptation more broadly.

Beyond recalibration: selective adaptation

Next we apply the ideal adapter framework to a phenomenon known as selective adaptation (Eimas & Corbit, 1973; Samuel, 1986). Traditionally, selective adaptation is thought to be due to mechanisms that are distinct from those underlying perceptual recalibration. We will show, however, that the cumulative build-up of selective adaptation is captured by the same belief-updating model introduced in the previous sections.

Selective adaptation occurs after repeated exposure to a single phonetic category, and is characterized by a *negative* aftereffect, where fewer items on a phonetic continuum are classified as the adapted category. For instance, and as we will discuss in more detail below, Vroomen et al. (2007) found that repeated exposure to a *prototypical* /b/ audio-visual adaptor constructed from the /b/ endpoint of their /b/-/d/ continuum (rather than the ambiguous midpoint) resulted in *fewer* /b/ responses during test trials (and vice-versa for /d/).

Selective adaptation is broadly considered to be the result of either habituation of “feature detectors” which are sensitive to linguistically-relevant features of the acoustic speech signal, or contrast effects at a categorical level (Samuel, 1986). It is also generally acknowledged that selective adaptation operates at a range of different levels. On the one hand, non-speech phonetic analogues (like isolated F2 and F3 formant transitions) can selectively adapt a place of articulation contrast, which suggests that selective adaptation operates on relatively low-level auditory processing (Samuel & Kat, 1996). On the other hand, selective adaptation has also been shown to generalize between acoustically different but phonetically similar continua, suggesting that it does not depend solely on acoustic overlap (Samuel & Kat, 1996; Sawusch, 1977).

Incremental selective adaptation—As is the case with recalibration, there is little work on how selective adaptation builds up incrementally. Again, one notable exception is Vroomen et al. (2007), who also investigated the build-up of selective adaptation at varying levels of exposure. Selective adaptation was induced by repeated exposure to a prototypical audio-visual adaptor, made using the same video as the ambiguous audio visual stimuli from the recalibration conditions paired with the corresponding category endpoint ($x = 1$ for /b/ and $x = 9$ for /d/). Other than this, the design and procedure was exactly the same as the recalibration condition described above: listeners heard a total of 256 repetitions of one of these adaptors, and were tested on the same audio-only classification test after 1, 2, 4, 8, 16, 32, 64, 128, and 256 cumulative adaptor exposures.

Figure 8 shows the results for the first 64 cumulative exposures in the first exposure block for selective adaptation from Vroomen et al. (2007). Like with the recalibration conditions, selective adaptation builds up incrementally over the first 64 exposures (although there is quite a bit of noise due to the small number of observations). Can a belief updating model account for this data?

Qualitatively, the answer is yes, as shown schematically in Figure 9. Just like the recalibration condition, the distribution of cues that listeners encounter in the selective adaptation condition is unusual: the exact same cue value is repeated over and over again.[14] In natural speech there is inevitable variability in the cue values used to realize a single phonetic category, and thus the level of consistency in this experiment is highly unusual. A belief updating model predicts that listeners will adjust their beliefs about the variance of the adapted category as a result of these unusual statistics. The results of adapting to low-variance exposure to, e.g., a prototypical /b/ is that the /b/ category 'shrinks', and cues values that were previously ambiguous become less likely under /b/ and thus more likely to be classified as /d/ (Figure 9, top). This is the negative aftereffect that characterizes selective adaptation.

While few studies have investigated the effect of unusual category variance on subsequent perception, one notable exception is Clayards et al. (2008). They showed that listeners who are exposed to /b/ and /p/ sounds whose VOTs have low variance show a sharper category boundary than listeners exposed to high variance distributions. Moreover, the difference in category boundary slopes is exactly as predicted by the category variances, via an ideal listener model. This suggests that listeners adjust their categorization behavior based on recently experienced within-category *variance* of acoustic cues, exactly as the ideal adapter framework predicts.

Modeling the build-up of selective adaptation—The ideal adapter framework—and the model introduced above—makes further predictions about how classification behavior depends on the *amount* of exposure, as well as its statistical properties. In order to evaluate the ability of belief updating to quantitatively account for the build up of selective adaptation, we fit the same model that was fit to the recalibration data above to the first 64 exposures of the selective adaptation data from the first block of each participant in Vroomen et al. (2007).

Results—The model also fits the selective adaptation data well ($r^2 = 0.83$), as shown in Figure 10. As the model predicts, selective adaptation starts very weakly, gradually becoming stronger with further exposure. The build-up of selective adaptation also appears to accelerate, unlike recalibration, with further exposure producing even larger increases in the strength of the selective adaptation effect, which is also captured by the model. Thus, while selective adaptation is not usually considered the result of belief updating or distributional learning, the same belief updating model which describes well how listeners

[14]There is of course additional perceptual uncertainty or noise variance (Feldman et al., 2009), but the combined variance from sensory uncertainty and actual stimulus variance across trials is still less for an identical repeated stimulus than would be expected with normal levels of variability.

integrate each observation with their prior beliefs during recalibration *also* describes the build-up of selective adaptation.

The best-fitting (MAP estimate) hyperparameters for the selective adaptation data are also nearly identical to the recalibration estimates: the effective prior sample size for the variances is $\nu_0 = 64$ (vs. 71), and for the means, $\kappa_0 = 13$ (vs. 11). The visual cue weight is $w = 0.51$ (vs. 0.53). This demonstrates two points. First, as with recalibration, these values correspond to very small effective prior sample sizes, which suggests that listeners do not believe that their vast amount of prior experience with /b/s and /d/s is very relevant in this situation.

Second, even though these parameter values correspond to higher prior confidence in the variance than the mean, the model nevertheless accounts for selective adaptation via shrinking category variance. This is because, unlike the ambiguous adaptor, for the prototypical adaptor there is no difference between the observed and expected category means. In this case, the updated estimate of the variance is simply an average of the prior expected and the observed variance, weighted by the effective prior sample size and actual sample size, respectively (see Appendix A, Equation (32)). For the best-fitting estimate of the effective prior sample size $\nu_0 = 64$ and an observed variance of zero, this means that the model believes the category variance to have halved after 64 exposures.

Discussion

These results suggest that, at least for the classification results of Vroomen et al. (2007), it is not necessary to invoke a qualitatively distinct process to explain selective adaptation. The ideal adapter also makes the correct qualitative predictions for a variety of other selective adaptation experiments (e.g., Eimas & Corbit, 1973; Miller, Connine, Schermer, & Kluender, 1983). It might thus be tempting to assume that *all* selective adaptation effects can be reduced to distributional learning of the type proposed here. Such an account will face several serious challenges. This includes the need to account for complex reaction time effects induced by different types of selective adaptation (Samuel, 1986; Samuel & Kat, 1996), about which our ideal adapter framework does not yet have anything principled to say. Rather than discuss this and other challenges to an ideal adapter account of selective adaptation here, we merely note that much work remains to be done in fleshing out the predictions of the ideal adapter framework for selective adaptation more generally.

In particular, three directions for future work stand out. First, in order to link the ideal adapter framework to reaction time data, process models have to be developed. Second, our account of selective adaptation depends on the listener's *perceived* variance of each distribution, which may not be the same as the measured variance of some physical cue. We have made the simplifying assumption here that the perceived variance is the same as the actual variance. For extreme cases, like the case of selective adaptation studied here, where there is *no*(or minimal) variance in the physical signal, this assumption leads to implausible asymptotic behavior: with more and more exposure, the category should shrink to nothing, leading to an asymptotic aftereffect of −1. Of course, between 128 and 256 exposures, the negative aftereffect induced by selective adaptation appears to continue to grow stronger (Figure 8), suggesting that it has not yet reached its actual asymptote, but we do not know

what that is for this paradigm. The distinction between perceived and actual variance becomes even more important when making quantitative predictions about different levels of variance. If the sensory uncertainty (or noisiness in the perceptual system) contributes substantially to the perceived variance (for preliminary evidence, see Clayards et al., 2008; Feldman et al., 2009; Kronrod, Coppess, & Feldman, 2012), this is expected to reduce or eliminate predicted effects of variation in physical variance.

Third, our account opens the door to alternative interpretations of particularly challenging aspects of listeners' classification behavior after selective adaptation (Kleinschmidt & Jaeger, 2013). More generally, recent work on non-linguistic sensory adaptation (mostly in low-level vision and audition) has revealed that many negative aftereffects which were originally attributed to the fatigue of neuronal feature detectors are better explained by neural populations adjusting their processing to maximize the transmission of information about the current stimulus ensemble (Brenner, Bialek, & de Ruyter Van Steveninck, 2000; Fairhall et al., 2001; Gutfreund, 2012; Kohn, 2007; Sharpee et al., 2006; Webster, Werner, & Field, 2005), which parallels recent developments in the understanding of perceptual learning in low-level perceptual tasks (Bejjanki, Beck, Lu, & Pouget, 2011; Harris, Gliksberg, & Sagi, 2012). Maximizing information transmission depends on the statistics of the environment at many different levels.

For speech perception, sometimes the relevant statistics are at the level of the phonetic generative model—the cue distributions for each category—but sometimes they are at a different level, such as the distributions of categories themselves or the spectral characteristics of background noise which must be ignored. For instance, J. Huang and Holt (2012) found that the classification boundary between "bet" and "but" could be manipulated simply by preceding exposure to a pure tone: when the frequency was near the second formant frequency of "bet", listeners made fewer "bet" responses. This seems incompatible with the idea that selective adaptation is due to the listener updating their beliefs about the distribution of cues associated with a particular category, but it is entirely consistent with a more general view that adaptation reflects changes in the listener's beliefs about which cues are more likely to occur—or be behaviorally relevant—across different levels of processing. That is, even though selective adaptation may not always represent updating beliefs about the talker's generative model, it may still serve the same purpose: efficient processing of linguistically-relevant acoustic signals in a world where the statistical properties of those signals vary across situations (Kleinschmidt & Jaeger, 2013). However, straightforward application of a model like the one presented here to perceptual inference at lower levels has not been successful (Stocker & Simoncelli, 2006) and more work remains to flesh out this connection.

For the current purpose, it is sufficient to conclude that not everything that looks like selective adaptation requires an explanation in terms of a separate computation. Despite the fact that recalibration and selective adaptation are typically considered to be qualitatively distinct phenomena, we have shown that a single belief updating model can account for the early, incremental build-up of both of these effects. This is achieved under essentially identical assumptions about prior effective sample sizes and the audio-visual cue weight (the hyperparameters in our model). We have also shown that the process of the listener updating

their beliefs about category variance provides a likely explanation for at least some adaptive behaviors during speech processing (see also Clayards et al., 2008). In the next section, we explore the consequences of this shrinking category variance in response to a repeated adaptor stimulus for prolonged exposure to a repeated stimulus.

Effects of prolonged repeated exposure to the same stimulus

The ideal adapter framework predicts that when presented with a single, repeated stimulus, the listener should shrink the variance of the repeated category. This leads to predictions about the incremental effects of selective adaptation by a prototypical sound (discussed in the previous section), but it also makes an interesting prediction about repeated exposure to an *ambiguous* sound. Even though listeners typically do not encounter a physically identical sound repeatedly in real life, the use of repeated sounds is common in perceptual recalibration experiments, which often use a single repeated token (e.g., Bertelson et al., 2003; Vroomen et al., 2004) or multiple words where the critical segment is replaced with the same ambiguous sound (e.g., Norris et al., 2003; Samuel, 2001; but see Kraljic & Samuel, 2005; Reinisch & Holt, 2014).

Recall that according to the belief-updating model presented above, in the recalibration condition of Vroomen et al. (2007) listeners do not perceive the adaptor cue value as fully ambiguous. Rather, their perceived cue value combines the ambiguous acoustic cue value and the prototypical visual cue value with roughly equal weight (visual cue weight parameter $w = 0.53$). The belief-updating model predicts that with repeated exposure to this not-quite-ambiguous adaptor, after the shift in the mean to the observed cue value, the low observed variance will eventually lead to the category shrinking and pulling back from the middle of the continuum. This prediction is illustrated in Figure 11.

This results in an eventual *decrease* in the likelihood that the adapted category assigns to the auditorily ambiguous test stimuli in the middle of the continuum and predicts that the positive after-effect associated with recalibration will eventually weaken and possibly even reverse if the category pulls back far enough. We have seen a hint that this may in fact occur: recalibration seems to effectively level off by 64 cumulative exposures in the data from Vroomen et al. (2007). In this section, we further test this prediction by looking at the additional data that Vroomen et al. (2007) collected for up to 256 cumulative exposures.

Data

As discussed above, Vroomen et al. (2007) exposed listeners to a total of 256 repetitions of the audiovisual adaptor. Figure 12 shows the results from both conditions, including the test trials at 128 and 256 cumulative exposures. As qualitatively predicted, the negative aftereffect associated with selective adaptation grows stronger with further exposure, while the positive aftereffect associated with recalibration plateaus and even begin to decline after 256 exposures.[15]

[15]While the decline in recalibration appears to be relatively modest in the figure here, Vroomen et al. (2007) actually found much stronger declines in recalibration in their full data set, where recalibration essentially disappears by 256 exposures. Our replication presented in the next section establishes that such a decline can occur in the first block (Figure 15), suggesting that the failure to observe it strongly in the first block of Vroomen et al. (2007) is due to individual differences, which are substantial.

Results and discussion

To quantify these predictions and test whether the belief updating model can account for the decrease in recalibration, we fit it to the data from all 256 exposures to the ambiguous and prototypical adaptors. The model fits well when fit to either the ambiguous or prototypical conditions individually ($r^2 = 0.86$ in both cases). To further test the model's ability to provide a unified explanation of recalibration and selective adaptation, we fit the model to both conditions simultaneously.

Figure 13 shows that the model simultaneously fits the behavioral data in both conditions quite well ($r^2 = 0.93$ overall, and $r^2 = 0.86$ and 0.85 for the ambiguous and prototypical subsets, respectively). There is no loss of goodness-of-fit from fitting both conditions simultaneously. The best fitting hyperparameters were $\nu_0 = 100$, $\kappa_0 = 17$, and $w = 0.47$. These hyperparameters are very similar to the best-fitting hyperparameters for the first 64 exposures only, as well as the estimates when each condition is fit separately.

In particular, the best fitting models for the initial build-up (first 64 exposures) and prolonged exposure (256 exposures) place roughly the same weight on the visual cue ($w = 0.53$ vs. $w = 0.47$, respectively, where w could range from 0 to 1), even though the decay in recalibration is virtually absent during the first 64 exposures. That is, just based on the rate at which recalibration initially accumulates, the model predicts the later decay without further assumptions.

Exploring the predictive power of the ideal adapter

One of the benefits of a model like we present here is that it makes quantitative predictions which go beyond existing data. In this section, we explore the ideal adapter framework's predictive power. The fact that the ambiguous and prototypical conditions from Vroomen et al. (2007) can be accounted for by a single set of belief updating model hyperparameters suggests that the recalibration and selective adaptation effects in this experiment are not qualitatively distinct but rather endpoints on a continuum of adaptation effects. The ideal adapter framework predicts that for observed cues which are less ambiguous, the low variance of the adaptor distribution will be detectable with fewer observations, causing an earlier (and lower) peak in the positive aftereffect detected in the recalibration condition.

In order to test this prediction, we replicate and extend the design of Vroomen et al. (2007), adding intermediate conditions where the acoustic component of the audio-visual adaptor is neither fully ambiguous nor fully prototypical, but somewhere in between. As a strong test of the predictive power of the model, we ask whether hyperparameters fit to the ambiguous/recalibration and prototypical/selective adaptation conditions in our replication can be used to predict adaptation behavior in new situations which have not been studied before.

Methods

We developed a novel web-based paradigm to efficiently collect phonetic categorization data from a large number of participants, adhering as closely as possible to the methods of Vroomen et al. (2007). In addition to the ambiguous and prototypical conditions of Vroomen

et al. (2007), we added two intermediate conditions: intermediate-ambiguous, and intermediate-prototypical.

Stimuli—Stimuli were identical to those used by Vroomen et al. (2007), who generously shared their materials. The audio stimuli were items from a nine-item synthetic /aba/-/ada/ continuum, created by shifting F2 locus in equal mel steps from a prototypical /aba/ value to a prototypical /ada/ value, holding other parameters constant (Vroomen et al., 2004). The visual stimuli were natural videos of a male talker articulating /aba/ and /ada/.

Audio-visual adaptors for the ambiguous and prototypical conditions were constructed as in Vroomen et al. (2007), by matching the video with the participant's most ambiguous continuum item x_{bd} or the corresponding continuum endpoint, respectively. For the two intermediate conditions, the audio component was offset by one (intermediate-ambiguous) or two (intermediate-prototypical) positions towards the video category endpoint (Figure 14).

Procedure—Participants first performed the same pre-test /b/-/d/ classification task as in Vroomen et al. (2007) (described above). After the calibration phase, participants were split into four exposure conditions (Figure 14): ambiguous and prototypical conditions as in Vroomen et al. (2007), plus two intermediate conditions. For the intermediate-ambiguous condition, the adaptors were constructed from the continuum item one position over from the most ambiguous position, in the direction of the endpoint corresponding to the video category. The intermediate-prototypical adaptors were constructed from the continuum item two positions over from the most ambiguous item. Each participant did two exposure blocks, one /b/ and one /d/.

Each block consisted of a total of 128 audio-visual exposure trials. Audio-only /b/-/d/ test trials were interspersed after 1, 2, 4, 8, 16, 32, 64, 96, and 128 cumulative exposures, in blocks of 6 or 12 trials (2 or 4 repetitions of each participant's 3 most ambiguous stimuli). Following the first exposure block, participants took a short break and completed the second block, with an audio-visual exposure stimulus from the opposite category but same condition.

The experiment was conducted over the web, via Amazon's Mechanical Turk crowd-sourcing service using a custom Javascript application.[16] Response keys and block order were counter-balanced across participants. The experiment took no longer than 45 minutes to complete.

Participants—A total of 280 participants were recruited. Since participants were being run remotely, using a variety of audio equipment to complete the experiment, a number of quality checks were required. First, our task was only made available to workers whose location was listed as the US and had more than 95% of their previous work accepted for payment (an automatic quality control measure offered by Amazon). Second, based on the calibration task, participants with unusual category boundaries (most ambiguous stimulus

[16]This source code and a working demo are available from http://hlplab.wordpress.com/2013/09/22/phonetics-online/

not one of the middle three positions 4, 5, or 6 found by Vroomen et al., 2007) were automatically excluded from the remainder of the experiment. 60 participants were excluded for this reason. Third, participants who classified the two endpoint stimuli and their nearest neighbors (x = 1, 2, 8, or 9) with less than 70% accuracy were also excluded. 25 additional participants were excluded for this reason.

Fourth, in order to ensure that participants were actually watching the videos during the exposure phase, catch trials were interspersed throughout exposure (as in Vroomen et al., 2007). These trials were identical to normal audio-visual exposure trials, except for a small white dot which flashed for one frame above the talker's lip. On these trials, participants were instructed to press the space bar to indicate they saw the dot. Participants who missed more than a total of 20% or more than 50% on any one block were excluded from analysis and replaced (13 participants). After this exclusion, the overall catch trial accuracy rate was 96% (compared to 93% reported by Vroomen et al., 2007).

Following good statistical procedure (Simmons, Nelson, & Simonsohn, 2011), these exclusion criteria were fixed before beginning data collection and automatically executed by our experiment software. Data from a total of 182 participants remained for analysis.

Modeling—We fit the belief updating model introduced above against only the ambiguous and prototypical conditions. We chose to fit these conditions for two reasons. First, we want to replicate the model fits to the data from Vroomen et al. (2007) on a novel language (and in our novel paradigm). Second, we want to test the ability of the model predict the effect of cumulative exposure in the two novel, intermediate conditions, based on conditions which have already been investigated. The posterior distribution of model hyperparameters (ν_0, κ_0, and w) given the ambiguous and prototypical conditions can be used to generate predictions for the intermediate conditions (by plugging in the adaptor values, offsets of one and two from the most ambiguous stimulus, $x_{bd} \pm 1$ and $x_{bd} \pm 2$, for the intermediate-ambiguous and intermediate-prototypical conditions, respectively, with the sign determined by the visual component).

Results

Here we focus on the evaluation of the incremental belief updating model developed in the previous sections. Additional data analyses confirmed that our web-based design replicates the results of Vroomen et al. (2007) and further supports our interpretation of the data. These analyses can be found in the Supplementary Material to this article.[17] Figure 15 shows the results from the four conditions, along with the predictions from the model fit to the ambiguous and prototypical conditions. First, our results replicate those of Vroomen et al. (2007). Exposure to the prototypical adaptor leads to negative aftereffects which build up gradually and become stronger throughout exposure, while the exposure to the ambiguous adaptor leads to positive aftereffects which peak and eventually fade, becoming negative after 128 exposures (see Supplementary Material). In fact, the recalibration decline that we find is even stronger than that found by Vroomen et al. (2007), with 128 exposures being

[17]The Supplementary Material are available from http://www.bcs.rochester.edu/people/dkleinschmidt/pubs/KleinschmidtJaeger-SupplementaryMaterial.pdf

enough to almost completely erase any positive aftereffect. Vroomen et al. (2007) observed such a reversal after 256 exposures in their full data set, but only a slight dip in the strength of the recalibration effect by the end of the first block of exposure (Figure 12). Our belief updating model fits the data from the ambiguous and prototypical conditions quite well (r^2 = 0.91 overall, vs. r^2 = 0.93 when fit to Vroomen et al., 2007).

Second, these results validate the predictions of the belief updating model for intermediate conditions, both qualitatively and quantitatively. Qualitatively, as predicted, the two intermediate conditions reveal that the ambiguous and prototypical conditions are endpoints of a continuum of adaptation effects. This is visually clear from Figure 15 and also borne out by statistical analysis (see Supplementary Material). Quantitatively, the belief updating model hyperparameters that were fit *only* to the ambiguous and prototypical conditions accurately predict the adaptation build-up in the two intermediate conditions. In fact, the fit to the intermediate conditions is as good as the fit to the conditions the model was trained on (r^2 = 0.96 for the two intermediate conditions vs. r^2 = 0.91 for the ambiguous and prototypical). This is encouraging given that individual participants show substantial variability in their adaptation behavior and each condition consisted of an entirely different group of participants.

We can further compare the model predictions and the data by looking at the point at which the aftereffect crosses over from being positive (recalibration-like) to negative (selective-adaptation-like). In Figure 15, these cross-over points are indicated by arrows—black for the model predictions and colored for the data—and Figure 16 shows that the observed and model-predicted cross-over points are strongly correlated, even though the cross-over point itself was not explicitly fit to the data. Thus, the belief updating model very effectively predicts the point at which behavior switches from recalibration-like to selective-adaptation-like for intermediate adaptors of varying ambiguity, based solely on the fit to ambiguous and prototypical adaptor data.

The best-fitting (MAP estimate) effective prior sample sizes for the category variances and means were small: ν_0 = 88 and κ_0 = 2.9, respectively. This again closely resembles what we found for the Vroomen et al. (2007) data. It also corroborates the conclusion that participants in these studies consider their previous experiences with /b/ and /d/ to be not particularly relevant in these situations, and moreover replicates the finding that, for these stimuli and this experimental paradigm, listeners seem to place more confidence in their prior beliefs about the category variances than the means ($\nu_0 > \kappa_0$ both here and in all previous model fits). Finally, as with the data from Vroomen et al. (2007), the best-fitting cue combination weight had audio and visual cues weighted roughly equally, placing in this case slightly more weight on the visual component (w = 0.63).[18]

Discussion

These results demonstrate two points. First, by replicating the results of Vroomen et al. (2007), they show that web-based platforms are a viable way to investigate phonetic effects

[18]The slightly higher weight for visual cues in our data might be a consequence of using stimuli derived from a Dutch /b/-/d/ continuum, which are similar but presumably not identical to those typically experienced in American English.

such as recalibration and selective adaptation which depend on the particular acoustic parameters of the stimuli. Most participants were enthusiastic and engaged in the study, and using sensible exclusion (based on pre-test performance or catch trials during exposure) those who are not fully attentive or who cannot hear the stimuli properly can be excluded.

Second, the ideal adapter framework for understanding phonetic adaptation presented above can explain both the phonetic recalibration *and* selective adaptation data from Vroomen et al. (2007) based on the mean (ambiguous vs. prototypical) and variance (none) of the adaptor distributions. This framework predicts that intermediate adaptors should produce intermediate adaptation effects, and moreover, the formal, quantitative model based on these principles makes specific, quantitative predictions about the effect of cumulative exposure to these intermediate adaptors. Both these qualitative and quantitative predictions were borne out in the data presented here.

Concluding Part I: Adaptation as inference under uncertainty about the statistics of the generative model

In this first part of this article, we have formulated the ideal adaptor framework, building on previous work on speech perception within the ideal listener framework (Clayards et al., 2008; Feldman et al., 2009; Norris & McQueen, 2008; Sonderegger & Yu, 2010). We have applied the ideal adaptor framework to two phonetic adaptation phenomena commonly considered to be due to distinct mechanisms, perceptual recalibration and selective adaptation. To do so, we derived a Bayesian belief updating model from the ideal adaptor. This framework allowed us to formalize the intuition that perceptual recalibration is a form of distributional (statistical) learning that changes the underlying representations of phonetic categories (e.g., Norris et al., 2003; Kraljic & Samuel, 2005; Maye et al., 2008; Bertelson et al., 2003). The Bayesian belief updating model provides a good quantitative fit to perceptual recalibration data (r^2s > .8). The model also qualitatively captures the cumulative effect of exposure in perceptual recalibration experiments, including the previously observed reversal of initial effects after prolonged exposure to the same stimulus (Vroomen et al., 2007).

Going beyond perceptual recalibration, the model suggests selective adaptation, too, is at least in part due to distributional learning similar or identical to that observed during perceptual recalibration. In addition to good quantitative and qualitative fits to selective adaptation data, our Bayesian belief updating model predicted a continuum from selective adaptation to perceptual recalibration, which was indeed observed. This serves as a demonstration of the ability of the ideal adapter framework to make quantitative predictions that go beyond existing data.

It is worth mentioning that the *model* which is used to generate those predictions in this case makes many assumptions that are not intrinsic to the framework but rather are made for convenience, tractability, or to better illustrate the basic mechanics of belief updating (for discussion, see the Appendix). These assumptions are largely justified for this particular experimental paradigm (see Table A1 in the appendix). They must, however, be considered carefully when thinking about how this particular *model* might generalize to other paradigms. Support for the ideal adaptor framework as a general framework of adaptation

beyond the type of experimental paradigms considered here comes from research finding qualitatively similar patterns of speech adaptation in situations that more closely resemble the complexity of every day speech perception (Eisner & McQueen, 2006; Reinisch & Holt, 2014).

Novelty of the ideal adapter: Putting learning front and center

Many of the ideas we have discussed so far are anticipated in more or less explicit form in previous work on speech perception. Our approach to speech perception is novel in two ways. First, the ideal adapter puts learning front and center in speech perception, as an unavoidable consequence of the combination of the task of the speech perception system—mapping cues to categories—and the world where that task is carried out—where the cue-category mapping varies from one situation to the next. In this way of looking at the speech perception system, learning or adaptation is a necessary part of normal speech perception. Existing models of speech perception—Bayesian and otherwise—for the most part do not learn or adapt (Clayards et al., 2008; Feldman et al., 2009; McClelland & Elman, 1986; Norris, 1994; Norris, McQueen, & Cutler, 2000). This is, in general, not an in-principle limitation of these modeling frameworks, but rather reflects simplifying assumptions made for the sake of tractability in previous work.

There are only a few models of speech perception that learn,[19] all of which have been proposed in recent years based on findings that speech perception is highly flexible (Lancia & Winter, 2013; Mirman, McClelland, & Holt, 2006). These models are connectionist models that adjust feedforward weights from acoustic cues to phonetic categories using Hebbian mechanisms. We postpone further discussion of these models until Part III below, except to say that they have only been applied qualitatively to the asymptotic effects of phonetic recalibration (if they have been directly applied to phonetic recalibration at all). As such, it is not clear whether they can, like the belief updating model we present here, simultaneously account for selective adaptation as well, or the detailed effects of cumulative exposure on the size and direction of phonetic adaptation. Another class of models that are broadly compatible with the ideal adapter is episodic or exemplar models (e.g., Johnson, 1997b; Goldinger, 1998; Pierrehumbert, 2003), which learn the distribution of sounds corresponding to linguistic units implicitly by storing raw acoustic traces. However, as we will discuss at length in Part II, despite similar motivations, the ideal adapter framework differs in an important way, because it abstracts away from individual episodes, both at the level of phonetic categories (or some other sublexical level of representation) and at the level of individual talkers and groups of talkers.

A deeper problem with existing models—even those that learn—is that they don't address how listeners manage to balance stability and plasticity in speech perception. Rather, learning rate is treated as a free parameter which is tuned such that the model behaves in a reasonable way. A model of the speech perception system ultimately has to address how stability and plasticity are balanced, because there is a wealth of evidence that people manage to do this very well. While the belief updating *model* we have presented above has

[19]Leaving aside for the moment models of *acquisition*, which we will discuss below.

free parameters that control the learning rate (the prior confidence parameters) and which were fit to the data, the ideal adapter *framework* provides a more principled way to approach the stability-plasticity trade-off, which we turn to in the second, more speculative and forward-looking part of this paper.

Part II

Human speech perception is characterized by both extreme flexibility and stability. Listeners can rapidly adapt to a novel pronunciation while not losing the ability to efficiently comprehend standard pronunciations. Part II of this paper focuses on the balance between stability and flexibility, and in particular how—in the ideal adapter framework—it is related to the structure of the world that the speech perception system has to operate in. There are two main aspects to this structure.

First, the generative model—the statistical properties of speech sounds for each phonetic category—can be *different* from one situation to the next. If the listener's beliefs about these statistics are substantially different from the current situation's actual generative model, then comprehension can be slowed or even incorrect. Situation-by-situation differences in the generative model require that listeners continuously infer the appropriate generative model, combining their prior expectations with current observations. In Part I we illustrated how phonetic adaptation can be understood as incremental belief updating, focusing intentionally on novel situations where listeners' prior experience was not likely to be very informative, allowing the influence of each additional observation to be seen more clearly.

However, there is a second relevant aspect to the speech perception system: generative models do not vary *arbitrarily* across situations, but are rather tied to, for example, *who* is talking. This means that listeners can expect to encounter the same—or similar—generative models again and again (and some more than others). In Part II of this paper, we focus on these *similarities* across situations that a listener encounters in the world, and discuss how the ideal adapter framework links this structure—with both predictable and unpredictable variation in generative models—with the stability and flexibility of the speech perception system. While a lot of future work is required to flesh out a computational cognitive model, we will present a tentative outline of how the ideal adapter framework formalizes the link between the distribution of generative models in the world, listeners' prior expectations, and their behavior. The ideal adapter framework's basic predictions are qualitatively supported by the available behavioral evidence. At the same time, the ideal adapter framework identifies a number of questions that we consider particularly critical for future research.

Recognize the familiar

While we have assumed that adaptation to a novel talker does not start from scratch, but rather from some form of prior beliefs about phonetic distributions (literally, the prior in the Bayesian belief-updating model introduced in Part I), we have so far paid little attention to the properties of this prior knowledge. For the examples entertained in Part I, such as the speech a listener is exposed to in a perceptual learning experiment, the specific prior beliefs that the listener brings to the situation are of little consequence because they will quickly be overwhelmed by the input listeners receive from the novel talker. In everyday life, however,

listeners will not only encounter talkers that differ starkly from previously experienced talkers. Instead, some talkers will have been encountered before and others will resemble previously experienced talkers to varying degrees. This means that *different aspects* of previous experience bear more or less strongly on the current situation. How listeners determine which previous experiences are the most relevant in a given situation is one of the goals of the second part of this article.

Consider, for example, what happens when you walk into a room where your best friend is talking. You can recognize them based on a variety of things, perhaps including their face, distinctive clothes they happen to be wearing, the fact that they are standing in a room in their own house, or the timbre of their voice. Because every individual speaks slightly differently, your best friend sounds different from other talkers, and it would be beneficial to be able to just "swap in" their particular cue statistics based on your vast experience with their speech in the past. Anecdotally, this is what happens when we encounter a highly familiar talker, and indeed it is broadly accepted and empirically supported that we have such talker-specific representations, and use them online in speech perception (Creel et al., 2008; Creel & Bregman, 2011; Goldinger, 1996; Nygaard & Pisoni, 1998; Palmeri, Goldinger, & Pisoni, 1993; Remez, Fellowes, & Rubin, 1997), although questions remain about the limits of this ability (Magnuson & Nusbaum, 2007; Pardo & Remez, 2006).

Talker-specificity in speech perception reveals an important insight: knowledge about the statistics of speech cues needs to be *structured* to provide the full benefit to the listener. This has long been recognized by, for example, exemplar-based theories of speech perception, where storage of rich acoustic details of each episode leads naturally to persistent, talker-specific representations of the variability of speech sounds, either implicitly or explicitly (Goldinger, 1996, 1998; Johnson, 1997a, 2006; Pierrehumbert, 2003). In the language of the ideal adapter framework, the presence of talker-specificity means that, far from simply being engaged in continuous statistical learning, listeners are in fact using their previous experience to at the very least determine where to *start* such statistical learning in each situation. It is this second type of inference—inferences about what aspect of previous experience are most relevant to the current situation—that prevents listeners from having to re-adapt from scratch every time they encounter a talker (whether it is a novel talker or a previously encountered one). Although the existence of talker-specific knowledge is now broadly accepted, the consequences that the existence of such knowledge has for understanding speech perception is perhaps still under-appreciated. Specifically, most previously proposed models of speech perception, with the exception of certain exemplar-based approaches (Goldinger, 1998; Johnson, 1997a; Pierrehumbert, 2003), either do not learn (Clayards et al., 2008; Feldman et al., 2009; Norris & McQueen, 2008) or are what we will call 'flat' learners, without the ability to induce structure over talkers (Feldman, Griffiths, et al., 2013; Lancia & Winter, 2013; McMurray et al., 2009; Mirman et al., 2006; Vallabha et al., 2007). These models are insufficient to account for talker-specificity and related phenomena discussed below. Even looking beyond speech perception into the burgeoning literature on learning in the face of (latently) non-stationary statistics (e.g., Cho et al., 2002; Gallistel et al., 2001), most existing models cannot account for some of the

basic properties of speech perception discussed here.[20] We return to this point below, as it motivates the proposal we lay out in this second part of the article.

Generalize to the similar

There is also evidence that the structure of listeners' previous knowledge extends beyond the level of particular individual talkers. To take a somewhat extreme example, one of our colleagues moved from New York to Northern England, where a very different dialect of English is spoken. Initially, he had a great deal of difficulty understanding what anyone was saying, but after some number of months he found that it became easier to comprehend this accented speech, even when it was spoken by particular individuals he had never met before (Farmer, *personal communication*). In the language of the ideal adapter, we might say that our colleague learned something about how the cue statistics—or generative models—vary across individual talkers who share this accent, and that this gave him a head start in adapting to a new, but similarly-accented talker. Although somewhat less well-studied than talker-specificity, there is some evidence that people are broadly capable of such generalization, using their experience with groups of talkers to guide speech perception (Bradlow & Bent, 2008; Baese-berk et al., 2013; Creel & Bregman, 2011; Johnson, Strand, & D'Imperio, 1999; Johnson, 2006; Niedzielski, 1999; Sidaras et al., 2009). There are, in the world, a range of structures that group talkers, from common language community (leading to dialect and accent groups) to factors like gender or sexual orientation (B. Munson, 2007). The ideal adapter predicts that listeners should pick up on, and take advantage of, these groupings, to the extent that they are informative about the generative models of the corresponding talkers.

In order to take advantage of the structure of how generative models vary in the world, listeners need to *learn* this structure through experience. Listeners do not directly know how generative models are distributed in the world, just like they do not have direct access to the generative model behind each utterance they observe. This makes the problem of how to take advantage of any structure that might be there in the world another problem of inference under uncertainty. It is clear that listeners need to learn in order to group talkers together, since meeting two talkers with similar generative models could just be a coincidence, but meeting twenty such talkers likely indicates some underlying dialect group. However, this kind of higher-level learning is also required for talker-specificity. That is, in order to benefit from experience with a new talker in future occasions, listeners cannot just adapt to that talker's generative model for the current situation, but rather must, after sufficient exposure, remember what they have inferred about that particular talker's speech statistics so that they can deploy that knowledge on future encounters. The need for listeners to induce structure over their previous experiences with different generative models suggests that despite being rapid, adaptation—like that discussed in Part I—should persist because it reflects the listener's attempt to build a model of the current situation which can be useful in the future. This is another way of saying that phonetic adaptation is not just priming but a

[20]This also makes speech perception a potentially productive test domain for further development of general models of learning in a non-stationary world.

form of learning about the situation and/or talker (Eisner & McQueen, 2006; Kraljic & Samuel, 2005; Kraljic, Samuel, & Brennan, 2008, among others).

Outline for Part II

We will expand on these intuitions about how listeners can benefit from picking up on the structure of their experience with different cue statistics, and review the relevant literature. Our first goal is to argue that the ideal adapter framework provides a unifying view that ties together a range of previously stated intuitions and proposals about talker-specificity, generalization, and adaptation in speech perception. Our second goal is to lay out a preliminary formalization of the ideal adapter framework as it relates to how a listener could take advantage of structure in their prior experience. While we will not present any implemented, quantitative model simulations like in Part I, we believe that formalizing the framework has a number of advantages. First, it will highlight the computational parallelism between recognizing familiar talkers, generalizing to similar talkers, and what we've called adaptation to novel talkers in Part I, all of which reflect listeners' attempts to infer the appropriate generative model for the current situation. Second, having a formalization of a theory can sometimes help when our theoretical intuitions break down, such as when, for instance, speech perception is *not* talker-specific (e.g. Kraljic & Samuel, 2007). Third, formalizing can help guide future research. As will become clear below, the ideal adapter framework puts a major emphasis on statistics of the speech signal, and how those statistics differ across talkers and groups of talkers *in the world*. Relatively little research has looked at this, and the ideal adapter framework provides a potentially productive link between this kind of data and human behavior.

Part II is organized into four main sections, each of which addresses a different aspect of how the listener's need to infer the appropriate generative model in each situation results in different strategies based on the information available to them from structure in the world (via prior experience) and the speech they observe in that situation. First, we focus on how listeners can benefit from familiarity with a particular talker. Maintaining and using talker-specific beliefs about generative models allows listeners to forgo adaptation altogether when they encounter a familiar talker again. Second, we ask how listeners get to the point of having such talker-specific beliefs. Every familiar talker was once novel, and became familiar through experience. Combining the belief-updating logic that was the focus of Part I with talker-specific beliefs leads to the ideal adapter's prediction of talker-specific adaptation. Third, we turn to the question of where a listener *starts* when adapting to an unfamiliar talker. Even if a particular individual is unfamiliar, the listener often has experience with other similar talkers that can be informative. We will discuss evidence that listeners are indeed sensitive to this information, and how it can be formalized in the ideal adapter framework. Fourth, and finally, we tie all three strategies together, and discuss how connections between adaptation, recognition, and generalization are demonstrated by how listeners behave when they are not entirely sure what prior experience is more relevant.

Recognizing the familiar: Talker-specificity

We begin with situations in which stable beliefs about the generative model are maximally beneficial: particular familiar talkers. Listeners can benefit from experience with a familiar

talker to the extent that they consistently use a particular generative model—producing certain cue distributions—that is different from other talkers. By maintaining stable representations of that talker's particular generative model, the listener would be able to deploy those representations when the talker is encountered again in the future, removing the need to adapt to them again. If, as the ideal adapter predicts, listeners are taking advantage of this structure—that talkers tend to use the same generative model consistently—then listeners should generally process speech in a talker-specific way, and specifically should process speech from familiar talkers faster, more accurately, and more robustly.

What do we know about talker-specificity?

Talker-specificity is one of the best-studied and uncontroversial features of human speech perception. Talker-specificity is observed both in the form of the ability of listeners to explicitly identify particular individual talkers based on their speech, even when it is highly degraded (Bricker & Pruzansky, 1966; Palmeri et al., 1993; Remez et al., 1997; Sheffert, Pisoni, Fellowes, & Remez, 2002). Offline measures of speech processing also show talker-specificity. For instance, words are better remembered when they're spoken by the same talker at study and test (Goldinger, 1996; Palmeri et al., 1993). However listeners also use talker-specific information to benefit processing. Listeners are faster and more accurate at comprehending speech in noise when it was produced by familiar talkers, both for unaccented (Nygaard & Pisoni, 1998) and accented (Clarke & Garrett, 2004) talkers. These effects seem to operate at a low level in speech perception, with online measures like eye-tracking suggesting that listeners use experience with talker-specific productions as early as it is possible to detect (Creel et al., 2008; Mitterer & Reinisch, 2013).

How does the ideal adapter formalize talker-specificity?

We can formalize the intuition that listeners use talker-specific information to guide speech perception within a Bayesian inference framework in the following way (visualized in Figure 17). Speech perception is treated as an inference process, where the listener is trying to infer the talker's intended category *c*(or message, more generally), given some acoustic observation *x*. As discussed in Part I, this relationship is probabilistic, due to unavoidable variability in production and sensory uncertainty, and we can express the cue-to-category inference using Bayes' rule, as in Equation (1), as a combination of how likely a category is *a priori*(the prior), and how well it predicts the observation (the likelihood). Talker-specific information can come into play in both the prior and the likelihood. If a talker is more likely to produce /s/ (overall or because of lexical preferences), this should be taken into account in the prior probability assigned to /s/. More importantly, given that each talker might produce a different distribution of cues for /s/, the *likelihood* of any given observation also depends on the talker.[21] Treating both the likelihood and prior as conditional on the talker (denoted *t*) results, after applying Bayes rule, in a talker-specific posterior over possible intended categories:

$$p(c|x,t) \propto p(x|c,t)p(c|t) \quad (11)$$

[21]The belief-updating model implemented in Part I only models changes in the likelihood.

This is shown schematically for classifying a VOT value as /s/ or /ʃ/ in Figure 17: different talker-specific cue distributions (middle) result in different category boundaries (left).

However, this way of formalizing talker-specificity doesn't exactly capture the fact that the likelihood $p(x \mid c, t)$—the distribution of cues for each category—depends on the talker's *generative model*, rather than the talker's *identity* per se. We can represent a particular generative model with a vector of its parameters—things like the mean VOT for /b/, the variance of the frication frequency for /s/, etc.—jointly denoted as θ.[22] If the listener knew, exactly, the talker's actual generative model, then we could write the talker-conditional likelihood as directly conditional on that actual generative model:

$$p(x|c,t)=p(x|c,\theta=\theta_t) \quad (12)$$

Of course, we have argued above that listeners in general do *not* have direct access to the exact generative model, but rather some uncertain beliefs about it. That is, the listener's knowledge about the talker t 's generative model is better described as a probability distribution, $p(\theta \mid t)$, rather than a single value, θ_t(Figure 17, right). One consequence of this is that the talker-specific likelihood should, ideally, take into account this uncertainty by averaging the likelihood assigned by each possible generative model $p(x \mid c, \theta)$, weighted by how likely the listener thinks that particular generative model is for talker t, $p(\theta|t)$:[23]

$$p(x|c,t)=\int p(x|c,\theta)p(\theta|t)\mathrm{d}\theta \quad (13)$$

Making beliefs about the generative model parameters explicitly conditioned on talker identity introduces another level of structure above and beyond the belief-updating model in Part I. That model simply sought to infer the appropriate generative model parameters based on some, overall beliefs about what parameters were likely overall—the prior $p(\theta)$—and how well the current speech was explained by each possible generative model—the likelihood $p(x \mid \theta)$. The belief-updating model from Part I (implicitly) thus assumes that the same prior beliefs would be relevant, no matter the situation, and *all* that it can do to adapt to the current situation is to update its beliefs through distributional learning. This makes the incremental belief updating model and other flat learning models—such as simple recurrent networks, Chang et al., 2006; Elman, 1990, other connectionist models without further assumptions, Mirman et al., 2006; Lancia & Winter, 2013, non-hierarchical reinforcement learning, Gallistel et al., 2001, etc.—insufficient to account for some of the most basic properties of speech perception.

However, if listeners have prior beliefs that are specific to—conditioned on—a particular familiar talker, then they do not need to bring exactly the same beliefs to every situation.

[22]We represent the generative model via a list of its parameters as a notational convenience, and not to make any claims about whether or not the listener's actual generative model is a *parametric* model, where each category is, for instance, a normal distribution over its cues. Formally, we simply mean the parameters θ to be a description of the generative model, which might be as compressed as a vector listing the mean and variance of each category on each cue dimension, or as fine-grained as an infinitely long vector which just lists the actual likelihood assigned by each category to each possible sound.

[23]This further assumes that the listener recognizes talker t with complete certainty. In many cases, this is a reasonable simplifying assumption, but in many others it is not, and we will address the consequences of this in a later section.

Rather, by *recognizing* a familiar talker they can deploy the corresponding talker-specific prior beliefs to "adapt" to the current situation without needing to actually go through the process of distributional learning. In the next section, we will explore the question of how listeners get to the point of having beliefs about a particular talker's generative model, but first we discuss what we see as some of the most pressing issues for future work that this raises.

Open questions: talker-specificity

One of the consequences of thinking about talker-specific beliefs as distributions over generative models is that the inevitable uncertainty of these beliefs might lead to further uncertainty in how speech is classified, above and beyond the inherent uncertainty from the probabilistic nature of speech perception itself. As a result of having only uncertain beliefs about a talker's generative model, a truly ideal adapter would have more uncertainty about how to categorize a particular cue value when they are less certain about the talker's generative model. This could result in, for instance, shallower category boundaries than predicted by a true ideal listener with perfect knowledge of the generative model. We should stress here that maintaining uncertainty is not per se a bad thing. To the contrary, properly accounting for uncertainty about the talker's intended message is a strength of this approach, and it would be far worse to be overly-confident about a classification that turns out to be wrong. Under this view, because listeners should have less uncertainty about highly familiar talkers, classification behavior should be closer to what is predicted by the actual cue distributions. Alternatively, listeners might not take into account their full uncertainty in the generative model, and go with their best guess rather than averaging over the full distribution. Other intermediate approximations are possible, with the listener considering only a finite number of different possibilities. While processing of familiar talkers is overall faster and more accurate, we do not know how this is reflected in the detailed patterns of categorization behavior, which is required to effectively evaluate how much uncertainty listeners are maintaining about the current generative model. Working out the implications of these various ways of dealing with uncertainty is thus in our view a pressing topic for future work, both computational and behavioral.

Another question is whether talker-specific beliefs about the generative model are something that is unique to speech perception, or instead reflects more general perceptual strategies. That is, here we have focused on *talker*-specificity. However, the ideal adapter framework in principle predicts that listeners might form beliefs about any type of re-encountered situation/context, if these context/situations are reliably associated with systematically different speech statistics (generative models) and encountered sufficiently frequently. This could include beliefs about being in certain types of spaces (an echoey cathedral, a wide open field, a room with a light that buzzes at particular frequency, etc.). Of course, many of these properties are relevant to many other sorts of auditory processing, like sound source localization, and the use and deployment of context-specific beliefs about cue statistics might not be something that is unique to speech processing. Rather, they may reflect a general property of both higher- and lower-level sensory processing, so that the relevant inference might occur below the levels we have so far considered.

Learning talker-specificity through adaptation

So far, we've discussed the benefits of talker-specific knowledge (and how it can be formalized in the ideal adapter framework) assuming that the listener *is* familiar with some talkers and their particular cue statistics (generative models). But of course listeners start off knowing very little about a particular talker's generative model, and so in order to be able to make use of talker-specific cue statistics, listeners need to learn about those statistics from experience. In Part I, we discussed evidence that listeners rapidly adapt to cue statistics in a novel situation, as predicted by the ideal adapter faced with the lack of invariance. However, if a listener *only* adapted, then they would never be able to develop the sort of talker-specific representations that are now generally accepted to be a basic feature of the speech perception system. The ideal adapter framework thus also predicts that, when faced with the need to learn such talker-specific representations, listeners do not *just* adapt in the short term, but hold onto their updated beliefs about the generative model for whatever talker (or kind of situation, more generally) that they have adapted to.

What do we know about talker-specific adaptation?

Despite being very rapid, phonetic adaptation has been found to lead to stable, persistent representations. This suggests that rapid phonetic adaptation is, in many cases, better thought of as the result of listeners learning something about the talker they encounter in the experiment. A prime piece of evidence for this is that the effects of phonetic adaptation can remain strong even after intervening exposure to speech from another talker (Kraljic & Samuel, 2005, 2007), even after leaving the laboratory and returning 12 hours later (Eisner & McQueen, 2006).[24] Moreover, there's also some evidence that listeners can learn separate generative models for two different talkers in the same situation, even when utterances from the two talkers are mixed together (C. M. Munson, 2011). This would not be possible if listeners were simply tracking the short-run statistics of particular acoustic cues like VOT.

How does the ideal adapter formalize talker-specific adaptation?

In the ideal adapter framework, this sort of talker-specific adaptation follows naturally from the idea, introduced in the last section, that listeners are using talker-specific beliefs about the generative model. Talker-specific beliefs about the generative model are updated in a similar way as in the "flat" belief-updating model introduced in Part I: by bringing the cue statistics predicted by the generative model into better alignment with the observed cue statistics. The only difference is that instead of considering the statistics of *all* observations (as the belief-updating model of Part I implicitly does), the listener should consider only the statistics of observations produced by a particular talker in updating their beliefs about that talker's generative model.

Such talker-specific belief-updating is formalized in generally the same way as it was in the belief-updating model in Part I. For each observation x from talker t, the listener updates their uncertain beliefs about talker t 's generative model, $p(\theta| t)$, by combining them with

[24]While there are some situations in which intervening speech from another talker *does* disrupt adaptation (e.g. the /d/-/t/ condition in Kraljic & Samuel, 2007), this is broadly consistent with the more general prediction of the ideal adapter that listeners are tracking the overall distribution of generative models across situations, which we discuss below.

information about which generative models are more or less compatible with the observation x :[25]

$$p(\theta|t,x) \propto p(x|\theta)p(\theta|t) \quad (14)$$

Note that, as in Equation (13), the dependence on the talker is entirely driven by the talker-specific beliefs about the generative model; given a particular value of the generative model parameters, the talker's identity doesn't change the likelihood of an observation *x*. These updated beliefs can then be further updated with other observations, with the updated beliefs after the first observation x_1 from talker *t* serving as the starting point for updating beliefs after hearing another observation from talker *t*, x_2:

$$p(\theta|t,x_1,x_2) \propto p(x_2|\theta)p(\theta|t,x_1) \quad (15)$$

In this way, the ideal adapter can combine information from multiple encounters with talker *t*, even if they are separated by intervening speech from other talkers (Figure 18). Combining information from different encounters with a single talker is important for inferring accurate beliefs about that talker's generative model because these generative models are extremely complex, with many different phonetic categories, each of which is cued by multiple acoustic features whose statistics have to be tracked. Each observation thus only provides a little bit of information about the talker's whole generative model, and arriving at accurate beliefs with low uncertainty requires, in principle, quite a bit of evidence. While this point might explain why listeners continue to update talker-specific beliefs over multiple encounters (preliminary evidence for which is provided by C. M. Munson, 2011), it raises the question of how listeners achieve reasonably robust speech perception even for relatively unfamiliar talkers. Besides the speech they directly observe, listeners have another powerful source of information: the range of different generative models they have encountered in their past experience with other talkers. As we will discuss in the next section, this information—which we can formalize as the *prior* or *base distribution* over generative models—can narrow down the range of generative models that the listener needs to consider, and serve as a head start to belief updating.

Open questions: talker-specific adaptation

One question that talker-specific belief-updating raises is what the limits on talker-specificity in adaptation are, both in the short term and over the long term. In the short term, how many sets of beliefs can a listener simultaneously maintain and update? There is fairly good evidence that listeners can maintain at least two distinct sets of beliefs for novel talkers in the same context (C. M. Munson, 2011; Kraljic & Samuel, 2007). Talker-specificity has been observed in recognition memory for words for up to 10 talkers at a delay of one day (Goldinger, 1996). But for more subtle effects like sublexical recalibration—changes in category boundaries that generalize to words not already heard—it's not known how many

[25]Really, the listener's inference about the intended category depends on their inference about the generative model parameters, and vice versa; see Equation (5). Here we average over (marginalize out) possible category interpretations of *x* for simplicity's sake, as in Equation (7).

talkers can be tracked. We hope that future work will address how the number of talkers, and the overall similarity of their cue distributions, affects talker-specific adaptation.

It's also not known what the limits on talker-specific adaptation are in the long term. Do listeners *always* create a distinct, persistent set of beliefs for each new talker? If the listener has good reason to think that they will not meet the talker again, then it may make sense to adapt and then forget. But this would be difficult to measure in a laboratory experiment, because re-testing listeners after any appreciable delay requires a degree of logistical coordination that provides pretty good evidence that persistent representations will be useful. Moreover, it may in some cases be more efficient to group multiple, highly similar talker together, which is addressed in the following section.

Generalizing across talkers

An obvious question is where this belief updating process *starts* for a novel talker. In the language of the ideal adapter, what are the listener's prior beliefs when first encountering an unfamiliar talker? Before hearing any of a novel talker's speech, the only thing that the listener has to go on is their previous experience with *other* talkers. To the extent that there's structure to this experience that can benefit adaptation to a novel talker, it behooves the listener to take advantage of it, by *generalizing* their experience with similar talkers. At the highest level, all talkers of the same language are by definition similar, and thus one way in which experience with different talkers can help listeners is that it provides some information about the overall range, or distribution, of generative models that exist in the world, and hence that the listener should be prepared to expect from new talkers in the future. By picking up on the structure of previous experience in this way, listeners can get a head start in adapting to a new talker, because their previous experience can tell them that some types of generative models are overall more likely than others. For instance, across talkers of American English, a particular talker's mean /s/ frication frequency centroid (an important cue in distinguishing fricatives) generally falls in the range of 5 kHz to 7 kHz, and the mean /ʃ/ frication centroid typically falls in the range of 4 kHz to 6 kHz (Newman et al., 2001). This means that a listener can (probabilistically) rule out generative models where /s/ or /ʃ/ have mean frication centroid outside these ranges, which substantially narrows down the range of generative models that they have to initially consider, providing a head start to adaptation (Figure 19). There are further regularities at more specific ways of grouping talkers. For instance, female talkers tend to produce both /s/ and /ʃ/ with relatively high frication frequencies, while males tend to produce them with relatively low frequencies. Thus, knowing something about indexical variables—who is talking, and what kind of person they are—makes some of the variability in generative models predictable.

Moreover, there are regularities across talkers in the relationships *between* generative model parameters that a listener can take advantage of to give a further head start to adaptation. For instance, even though the range of mean centroids for /s/ and /ʃ/ overlap across talkers, *within* a talker the mean centroid frequency for /s/ is almost always higher than for /ʃ/ (Newman et al., 2001). Based on this information listeners can also (again probabilistically) rule out generative models where the relative mean centroid of /s/ and /ʃ/ is reversed from

the typical pattern. This sort of structure essentially cuts in half the number of generative models that need to be considered, a priori (Figure 19).

Constraints like these make the problem of adapting to a new talker, in principle, vastly easier than learning the language in the first place. The overall space of generative models that are likely to occur for a particular language or group of talkers is much more restricted than the space of generative models that occur across *all* talkers of all languages. Some of these constraints might be innate (and common, at least probabilistically, to the world's languages), but other obviously have to be learned for each particular language (like the range of mean VOTs allowed for each category Lisker & Abramson, 1964).

What do we know about how listeners generalize across talkers?

Generalizing across all talkers—First, there is some tentative evidence that listeners are sensitive to the overall range of generative model they have encountered. If listeners are sensitive to this range, then listeners who have experience with a broader range of generative models should, on average, be better prepared to adapt to unusual speech. Baese-berk et al. (2013) found that, as predicted, listeners who had to transcribe sentences from four talkers with four *different* foreign accents were more accurate on a fifth talker, with a fifth accent, relative to listeners who only heard a single accent (distinct from the test talker's).

Conversely, listeners also use the range of generative models they have previously encountered to narrow down the hypotheses they consider for a new talker. It follows that it should be *harder* to adapt to talkers whose generative models fall outside the typical range. This prediction is borne out in recent work. For instance, Idemaru and Holt (2011) exposed listeners to various combinations of two cues to a voicing contrast (e.g. /b/ vs. /p/), VOT and fundamental frequency (f0). Canonically, these two cues are positively correlated within a talker: higher VOTs occur with higher f0s, and correspond to voiceless stops like /p/. While uncorrelated VOT and f0 are rarer, they are also observed. Crucially, anti-correlated VOT and f0 are generally *not* observed. If listeners have implicit beliefs that reflect these correlations, this should make it harder to adapt to unnatural talkers for which the two cues are anti-correlated. This is indeed what Idemaru and Holt (2011) found: listeners were able to adapt to a two-dimensional distribution of VOT and f0 where f0 was uncorrelated with VOT. However, listeners were not able to fully adapt to a distribution where VOT and f0 were anti-correlated. Similarly, Sumner (2011) found that American English listeners were not able to adapt to a talker who always produced voiced stops (i.e. /b/, /d/, /g/) with substantial prevoicing (negative VOTs). Given that American English talkers typically produce these sounds with VOTs of 0 ms, this inability to adapt, too, might be a consequence of cross-talker generalizations based on prior beliefs.

Generalizing based on social group membership—Listeners often know (or can infer) more about a talker than that they are a speaker of English (or whichever language). To the extent that different a particular group of talkers systematically differs in their generative models from talkers in general, the listener can benefit by identifying whether a talker is a member of this group, and using their previous experience with this group to provide an even bigger head start to adaptation. For instance, there are dramatic differences

in how men and women produce many phonetic categories (e.g., Hillenbrand et al., 1995; Jongman, Wayland, & Wong, 2000; McMurray & Jongman, 2011; Newman et al., 2001), and thus the listener might (probabilistically) rule out a range possible generative models, further facilitating fast adaptation.[26]

One straightforward prediction of this is that in the absence of enough direct experience with a talker to directly converge on that talker's generative model, the listener's beliefs will be a combination of their (gender-specific) prior beliefs and whatever they manage to glean from what speech they have observed. For speech sounds that differ systematically between male and female talkers, such as fricatives, the listener's best guess about the appropriate generative model for the current speech is expected to depend on whether the listener believes the signal to stem from a male or a female talker. If, as we have argued above, listeners use these beliefs about the generative model to guide interpretation of speech sounds, then changing the perceived gender of the talker should change categorization (e.g., by displaying a picture of a male vs. female face along with the audio stimulus).

This prediction is born out in a study by Strand and Johnson (1996, see also ; Strand, 1999; B. Munson, 2011). Listeners heard sounds on an /s/-/ʃ/ ("sod"-"shod") continuum, made by varying the frication frequency from high (/s/-like) to low (/ʃ/-like), paired with either a male or a female face. Frication frequencies are typically lower overall for male talkers than for female talkers (Jongman et al., 2000; McMurray & Jongman, 2011; Newman et al., 2001), If listeners take this information into account then they should place the /s/-/ʃ/ boundary at a *higher* frequency if they think the talker is female, resulting in *fewer* /s/ responses. This was exactly what Strand and Johnson (1996) found. Analogous results have been found using vowels (Johnson et al., 1999) and using different vocal sources as cues to gender (rather than faces Strand & Johnson, 1996; Strand, 1999; Johnson et al., 1999; B. Munson, 2011).

Beyond gender, there is evidence that listeners also use their experience with the speech of different social groups as a source of information about where to start adaptation. In a vowel matching task where they were told that the talker was from Canada, listeners from Detroit, MI chose tokens with more Canadian raising as matching than when they were told the talker was from Detroit (Niedzielski, 1999). This suggests that listeners' perception of these vowels was biased towards what they expected to hear from a Canadian talker. There is, more generally, a growing literature on the extent to which perceived social group membership affects language comprehension (Drager, 2010; Hay, Warren, & Drager, 2006; Hay & Drager, 2010; Staum Casasanto, 2008; Sidaras and Nygaard, *submitted*).

How does the ideal adapter formalize generalization across talkers?

In the ideal adapter framework, the listener's beliefs about generative models are formalized as distributions over generative models, which assign more or less probability to each possible generative model. This is true both for beliefs about a particular talker's or

[26]Of course, as above, if the talker's generative model is actually outside the normal range for women, this would *hurt* adaptation. But, by definition, the distribution of the generative models for talkers in a particular group captures most of the talkers in that group, and so *on average* will be beneficial.

situation's generative model, like $p(\theta \mid \text{Susan})$. We can also formalize beliefs about what generative models are likely for any talker of the language—$p(\theta \mid \text{English})$—or for a female talker—$p(\theta \mid \text{female})$—or a Canadian talker—$p(\theta \mid \text{Canadian})$. When first encountering a novel talker, all the listener has to go on are these sort of beliefs about the distribution of generative models *across* individual talkers in some group.

Above, we have talked intuitively about how the listener's experience with the generative models of different talkers from a novel talker's language or social group can rule out a lot of generative models ahead of time and give a head start to adaptation. In the ideal adapter framework, this is a consequence of thinking about belief-updating as statistical inference, where the distributions over generative models corresponding to the listener's prior beliefs and the information from the speech they observe are combined according to Bayes rule. In order to converge on the particular generative model for the current situation, the listener has to allocate more and more probability mass to it. Consequently, the more probability is allocated to it ahead of time, the less information is required from observations.

The reason that more specific (or constraining) prior beliefs provide more of a head start is that by definition, probability distributions have to add up to one, and so spreading probability mass over a larger range of generative models means that less probability mass is allocated to each *particular* possible generative model. Thus, conversely, the more generative models that the listener can exclude a priori, the more probability they can allocate to each remaining generative model. If the actual generative model for the situation lies within this range, then fewer consistent observations are required in order to reach a particular level of probability for the actual current generative model.

The price that is paid for this head start is that more specific prior beliefs are less *flexible*. That is, if the prior beliefs are *wrong* and the actual generative model lies outside the range that has non-trivial prior probability, then much more evidence will be required to reach the same level of belief in the actual generative model.[27] This is why, in principle, the listener should only use specific beliefs when they are fairly certain that they are applicable. While this might seem like a problem for the ideal adapter framework, in the next section we will show that the ideal adapter framework also points towards a solution, which is to treat the problem of identifying which beliefs are applicable as another inference problem, which can be solved by combining top-down information (like visual recognition of a particular talker) with bottom-up information from the speech itself.

Open questions: Generalization across talkers

How do listeners know who to group together?—In order to make use of structure in the distribution of generative models at different levels—from talker-specific to groups of talkers to the overall base distribution—listeners have to somehow pick up on and *learn* this structure. There is some evidence that people can not only induce this sort of structure, but in some cases can do so surprisingly quickly. Bradlow and Bent (2008) had listeners transcribe sentences from accented talkers. In the *same talker* condition, the sentences were

[27]This is really a consequence of Bayes rule assigning posterior probability as the *product* of prior probability and likelihood. That means that what "adds up" is actually *log*-probability, which goes to negative-infinity as probability goes to zero.

all produced by a single accented talker, and listeners transcribed sentences from this same talker more accurately in a later test phase. In the *multiple talker* condition, listeners transcribed the same total amount of speech, but from four different talkers (distinct from the test talker). Despite the test talker being novel to them, listeners in the multiple talker condition showed the same benefit of experience as the same talker group. Listeners in a *single other talker* condition did not receive any benefit to exposure to a single accented talker when tested on the novel test talker, which suggests that listeners in the multiple talker condition have extracted some accent-level beliefs after a relatively small amount of exposure (on the order of an hour or two).

In the language of the ideal adapter framework, we might say that listeners need to impute structure to the variability in the generative models they have experienced. Much like Bayesian models of category learning during language acquisition (e.g. Feldman, Griffiths, et al., 2013; Perfors, Tenenbaum, & Regier, 2011), this structure induction can be thought of as another problem of inference under uncertainty. Each way of grouping previously encountered generative models is an imputation of some kind of structure, and each possible grouping may be more or less compatible with previous experience and a priori expectations about how likely particular sorts of groupings are. However, the ideal adapter framework, as articulated here, has very little specific to say about how this inference might work—even at a computational level—outside of very general predictions (e.g. simpler groupings are naturally preferred because of the Occam's razor property of Bayesian inference; Perfors et al., 2011; MacKay, 2003, pp. 343–356).

How are multiple, overlapping groups represented?—Each individual talker can belong to many different groups of talkers, all of which may be informative about their generative model. There are many ways that such a structure might be represented formally. One of the simplest possibilities is that talkers are nested within groups ("Mandarin-accented male talkers"), which are themselves nested within more general groups ("Mandarin-accented talkers"), and so on. This has the potentially problematic feature of not allowing generalization to a combination of groups that has not directly been observed. Another possibility is that groups might be represented, not as absolute generative model parameter distributions, but as offsets from the parameters predicted based on other factors (including other group memberships).[28] This would make it possible to generalize based on arbitrary combinations of features, without needing direct experience with a particular combination, but at the price of requiring the same offset for, say, male vs. female talkers regardless of their other group memberships.

These possibilities lead to at least two open questions. First, do listeners, in fact, generalize from experience with a male talker with a particular accent or dialect to a female talker of the same accent or dialect (and vice-versa)? If listeners can generalize within a gender but not across, this would suggest that they have beliefs that are specific to a particular combination of gender, dialect/accent, and possibly other indexical variables. Of course, it is possible that for *some* dialects/accents listeners have gender-specific beliefs, while for others they do not. This leads to the second question: do males and females differ in the same way

[28]For an analogous model applied to allophonic variation of phonetic categories, see Dillon, Dunbar, and Idsardi (2013)

across all dialects/accents of a language? If the ways that the generative models depend on these different types of indexical variables are all independent of each other, then it makes sense to separate them out into orthogonal features. However, gender differences themselves vary quite a bit across languages (Johnson, 2006), so it is not implausible that within a language there might be dialectal variation in gender differences as well. Moreover, it is altogether possible that for some groups, gender has the same (independent effect) while for others it does not. If this is the case, then it is in those kinds of groups where gender has an independent effect that the ideal adapter predicts that listeners are most likely to generalize experience across genders. The advent of the widespread availability and use of speech recognition technology presents an exciting opportunity to gather speech from large samples of individuals, potentially allowing some of these questions about the distribution of generative models in the world to be addressed.

Balancing stability and flexibility: When to adapt, generalize, or recognize

Thus far, we've addressed three strategies that listeners have for dealing with the lack of invariance: rapid adaptation to totally novel talkers, recognition of familiar talkers (and the resulting talker-specificity), and generalization across groups of similar talkers. In the ideal adapter framework, all three of these strategies are the result of the listener trying to infer the appropriate generative model with differing degrees of relevant prior experience. In this way of thinking, whether speech perception is flexible—as in adaptation to novel situations—or stable—as in recognition of familiar situations—depends on how confident the listener is in their prior beliefs, which in turn depends on how much relevant prior experience they have. In order to introduce the relationship between these strategies and the type and amount of relevant prior experience the listener has, we have so far addressed these strategies separately, assuming in each case that the listener *knows* —based on non-linguistic cues— what kind of prior experience is relevant to the current situation.

While such cues are often available, they are not *always* available, and sometimes they do not provide high certainty. Consider, for example, the case of running into someone that one remembers from somewhere but doesn't quite remember who it is. Sometimes visual or other cues are absent, such as when picking up the phone in the absence of caller ID, or listening to a conference call with many different people. In these cases where there is some uncertainty about *who* is talking, there is also uncertainty about what prior experience will be most helpful in figuring out that talker's generative model (and ultimately what they are trying to communicate). When top-down cues do not unambiguously identify the talker (or type of talker), the listener can still benefit from prior experience. Specifically, if the listener can identify a familiar talker (or the group that a talker belongs to) from their speech, listeners will have nearly the same benefits of prior experience as if they had known the talker's identity ahead of time.

The ideal adapter treats the problem of whether to be flexible or stable as inference under uncertainty at yet a third level, where the listener tries to infer what *type* of situation they are in, and hence what prior experience will be relevant in figuring out the generative model.[29]

[29]The other two levels being inferring the intended category, and inferring the situation's particular generative model.

There are, as always, two sources of information that can guide this inference: the listener's top-down expectations about what type of situation they are in (like visually identifying a talker's face), and bottom-up information from the speech signal (which may be more or less compatible with the speech predicted by their beliefs about each possible type of situation). The ideal adapter framework predicts that listeners use these two sources of evidence in order to balance stability and flexibility on the fly.

What do we know about how listeners balance stability and flexibility?— Balancing stability and plasticity requires that listeners deploy the most informative prior experience that is relevant in each situation. This is true whether or not they know *a priori* what prior experience is relevant. Thus, we argue, balancing stability and plasticity in speech perception requires that the listener dynamically combine top-down information about what kind of situation they are in with bottom-up information from the speech signal. However, little is known about how listeners trade off these two sources of information. What little we know suggests that listener do, in fact, use both bottom-up and top-down information to infer the type of situation they are in and what prior experience is currently relevant. Based on these inferences, listeners can then determine how flexible to be in the current situation. Evidence for this view is of two types, which we discuss in turn. First, listeners can recognize talkers that they are familiar with based on just bottom-up speech input. Second, listeners generalize experience with one talker to adaptation to another talker based both on the bottom-up similarity in their generative models and top-down expectations that generalization is appropriate.

Recognizing familiar talkers based on their speech: Listeners use information from the speech signal itself to help infer the talker's identity (and social group membership) in a way that shapes how they interpret that talker's speech. The speech signal includes non-phonetic features like (in English) f0 range, jitter, and shimmer that distinguishes talkers (Creel & Bregman, 2011; Pardo & Remez, 2006). Listeners use this information not only to recognize familiar individuals but also to infer social group membership. For instance, we discussed above evidence that listeners change their classification of vowels and fricatives based on the gender of a visually presented face. These same studies found that synthesized vowels and fricatives—with exactly the same phonetic cue values—were also classified differently depending on the vocal source wave that was used, with stereotypically female and male vocal sources having a similar effect as a visually-presented female face (higher frequency category boundaries). Interestingly, this effect is gradient, with the effects of voices rated as non-stereotypically male or female falling in between stereotypically male or female voices (Johnson et al., 1999; B. Munson, 2011; Strand & Johnson, 1996; Strand, 1999).

Beyond non-phonetic aspects of the speech signal, the fact that different talkers use different phonetic generative models means that *phonetic* aspects of the speech signal can also, in principle, provide information about talker identity or group. That is, the same talker- or group-specific beliefs about the phonetic generative model which benefit speech perception once a particular talker is identified can *also* conversely aid in identifying a talker. To the extent that they vary across talkers—and hence benefit from talker- or group-specific representations—phonetic features like formant frequencies and VOT are by definition

diagnostic of talker identity. Thus, the listener's knowledge about a familiar talker's cue statistics provides a rich source of information for identifying that particular talker based on their speech, which can in turn can affect how that same speech signal is mapped onto phonetic categories. Listeners do, in fact, take advantage of such phonetic information for talker recognition (Creel & Bregman, 2011; Pardo & Remez, 2006), and can explicitly recognize talkers with very high accuracy based on sine-wave (Remez et al., 1997) or noise vocoded speech (Sheffert et al., 2002). Such processing of speech removes all of the fine spectral detail (and much if not all of the voice quality information) but preserves enough phonetic information to allow for reasonably good comprehension.

Generalizing experience with one talker to another, or not: When there is uncertainty about what prior experience is relevant, information from the speech signal helps processing of novel talkers in two ways. First, in this paper we have generally simplified our discussion by focusing on a few phonetic features. However, in everyday speech perception the generative model is highly complex, covering many categories, each cued by many acoustic features. Somewhat paradoxically, when we consider that the listener must also infer the kind of prior experience that is relevant, this makes their job easier: even from a small amount of speech from a novel talker, listeners can in all likelihood recognize that none of their talker-specific representations provide a good match to the current situation. Second, even a small amount of information about the generative model can help identify whether, and which, more informative *group*-specific beliefs might be relevant, meaning that listeners have to fall back on the least informative beliefs only as a last resort. This is important because on the one hand, even a little bit of experience with one or more similar talkers provides a big boost to adaptation, but on the other hand, mistakenly assigning a talker to a particular group actually makes it harder to adapt to their actual generative model, as discussed in the last section. If listeners are in fact continuously trying to identify what, if any, familiar group a novel talker belongs to based on the speech they produce, then they should flexibly generalize from experience with one talker to another when it's warranted (by top-down expectations and bottom-up similarity) but adapt in a talker-specific way when it is not.

This is an important though comparatively understudied issue. A series of recent studies provides some tentative evidence for the predictions of the ideal adapter framework. These studies looked at how much listeners generalize perceptual learning for a male talker to a different, female talker (or vice versa; Eisner & McQueen, 2005; Kraljic & Samuel, 2005, 2006, 2007). In one version of these experiments listeners hear a male talker's production of a sound ambiguous between a /s/ and a /ʃ/, embedded in a word that indicates it was intended to be an /s/. After exposure, listeners are tested on two different /s/-/ʃ/ continua: one produced by the same male talker as exposure, and the other produced by a different (female) talker. In the case of /s/-/ʃ/ (and /s/-/f/), listeners show the expected recalibration effect—more /s/ responses after /s/ exposure—when tested on the same talker, but no effect when tested on the different, female talker (Eisner & McQueen, 2005; Kraljic & Samuel, 2005). However, in another version of the experiment using a /d/-/t/ contrast, listeners show the same recalibration effect on *both* the same- and different-talker test continua (Kraljic & Samuel, 2006). In such experiments, the only information that the listener has about the

unfamiliar female test talker comes from the test stimuli themselves. In the case of a voicing contrast like /d/-/t/, the overall distribution of the critical cue—VOT—is generally similar between the familiar male and unfamiliar female talker, and so the test stimuli provide information that these two talkers' generative models are overall quite similar, at least in terms of VOT and the voicing contrasts it cues. If listeners are using this information, then they should apply what they have learned about the male talker to the unfamiliar female talker, which would lead to the generalization observed by Kraljic and Samuel (2006). Conversely, the range of cues in the /s/-/ʃ/ male and female continua used by Kraljic and Samuel (2005) differ quite dramatically: the highest spectral centroid of the male talker's /s/ is still lower than the *lowest* /ʃ/ end of the female talker's continuum. Thus, the test stimuli from the female test talker provide evidence that these two talkers are quite different in how they produce this contrast, and thus it does not make sense to apply what has been learned about the male talker to the unfamiliar female test talker.

Kraljic and Samuel (2007) present evidence that listeners use both their top-down expectations about whether two talkers should be grouped and the bottom-up similarity of their cue statistics to guide not just processing of a totally unfamiliar test talker, but also to guide adaptation to multiple talkers. When exposed to both a male and female talker, whose ambiguous pronunciations are disambiguated in opposite ways (e.g. male as /d/ and female as /t/), listeners showed no overall learning effect between pre- and post-exposure classification of the two talker's /d/-/t/ continua, suggesting that they have tracked the statistics of these categories in a talker-independent way. But for the analogous procedure with /s/-/ʃ/, classification of each talker's continuum shifted in the typical way, suggesting that they have adapted talker-specific representations (Kraljic & Samuel, 2007).

This might be taken as evidence that listeners are simply unable or categorically unwilling to learn and maintain talker-specific beliefs about some phonetic categories. This broadly makes sense: for categories like /d/ versus /t/ which, in the listener's experience, show some variability across talkers but are not systematically produced differently by male versus female talkers, it is a reasonable guess that a male and a female talker encountered in the same context will produce them similarly and can be grouped together. Likewise, as noted by Kraljic and Samuel (2007), for categories that are generally produced differently by male and female talkers who are otherwise similar, it is a good bet that a male and female talker should *not* be grouped together.

However, if listeners are actually trying to *infer* whether or not two talkers should be grouped together—rather than simply relying on fixed heuristics—then enough of the right kind of experience which contradicts these biases should be able to overcome them in specific cases. For instance, even though listeners tend to generalize experience with VOT distributions from a male to female talker, with sufficient experience with a male and a female talker who have *different* VOT distributions, listeners do in fact learn talker-specific representations (C. M. Munson, 2011). Conversely, even though listeners tend to learn talker-specific representations of fricatives, by using a modified test continuum Reinisch and Holt (2014) found that listeners can generalize experience with a female talker's fricatives to a novel male talker.

How does the ideal adapter formalize the stability/plasticity trade-off?

A mixture of prior beliefs: The available evidence suggests that listeners can, on the fly, infer which type of prior experience is relevant for the current situation based on a combination of bottom-up speech information and top-down expectations. In the language of the ideal adapter framework, this in turn suggests that in situations where the listener is unsure about what type of prior experience is relevant, their prior expectations about what generative models they will encounter are a *mixture* of expectations given different types of prior experience. Each particular sort of prior experience can be thought of as a cluster of possible generative models, with more specific prior experience—like experience with a particular familiar talker—corresponding to highly concentrated, peaked clusters and more general experience—like with the range of generative models across all talkers in the language—being more spread out. How much each cluster contributes to the overall beliefs depends on how likely the listener thinks it is to apply in the current situation. How much the listener expects to encounter any *particular* generative model is thus a combination of how much it is expected given a particular set of beliefs, and how likely the listener thinks those beliefs are to apply.

One way to formalize this notion is based on representing the beliefs about generative models for a particular type of prior experience as a probability distribution over generative models. Again, for the sake of brevity, we refer to the generative model by its parameters[22] θ, conditional on the *type* of prior experience t : $p(\theta \mid t)$. Here, we use t to denote a more general form of talker identity, that covers beliefs ranging from particular individual talkers, like $p(\theta \mid t = \text{my father})$, to groups, like $p(\theta \mid t = \text{Boston accent})$, all the way to language-level groupings like $p(\theta \mid t = \text{English speaker})$. Above, we assumed that the listeners *knows* the type of prior experience which is relevant, and thus only a single conditional distribution $p(\theta \mid t)$ is relevant. When the listener is uncertain, though, multiple different conditional distributions might be relevant. The listener's beliefs about which type of prior experience is relevant can be formalized as a probability distribution over the types t —$p(t)$—and hence their overall beliefs about the generative model parameters as a *mixture* of the conditional beliefs, weighted by $p(t)$:

$$p(\theta) = \sum_t p(\theta|t)p(t) \quad (16)$$

Top-down information: Let's consider the case where the listener can recognize a familiar talker—call him Frank—with high certainty based on visual information. Then, the distribution over types of prior experience $p(t \mid \text{vis. info})$ is highly peaked:

$$p(t|\text{vis.info}) = \begin{cases} 1 & \text{if } t = \text{Frank} \\ 0 & \text{otherwise} \end{cases} \quad (17)$$

As a result, the listeners' beliefs about the generative model parameters, given the visual information that identifies the talker as Frank, is dominated by their beliefs about Frank's generative model:

$$p(\theta|\text{vis.info}) = \sum_t p(\theta|t)p(t|\text{vis.info}) \quad (18)$$

$$= p(\theta|t=\text{Frank})p(t=\text{Frank}|\text{vis.info}) + \sum_{t\neq\text{Frank}} p(\theta|t)p(t|\text{vis.info}) \quad (19)$$

$$= p(\theta|t=\text{Frank}) \times 1 + \sum_{t\neq\text{Frank}} p(\theta|t) \times 0 \quad (20)$$

$$= p(\theta|t=\text{Frank}) \quad (21)$$

That is, when the type of prior experience is known with reasonably high certainty from top-down cues, this mixture model reduces to the talker- or group-specific beliefs described in earlier sections in Part II. This formalization thus generalizes the notion of talker- or group-specific beliefs as conditional distributions. In this sort of mixture model, when top-down information about the type of situation is available, listeners can take advantage of it, which can in turn lead to effects of top-down information (like visually presented faces or other cues) changing the way that physically identical bottom-up acoustic signals are interpreted (Johnson et al., 1999; Hay & Drager, 2010; Staum Casasanto, 2008; Strand & Johnson, 1996). However, having a weighted mixture of different prior beliefs means that the listener does not have to rely *solely* on such top-down information to tell them which prior experience is relevant, but rather can also use bottom-up information.

Bottom-up information: Formalizing the listener's prior beliefs about generative models in this way provides a natural way of formalizing how bottom-up, phonetic information can help infer the type of prior experience that is relevant. Inferring the type of experience that is relevant comes down to inferring the type *t*, given some observations *x* and prior expectations $p(t)$, which we can think of as Bayesian inference at yet another level:

$$p(t|x) \propto p(x|t)p(t) \quad (22)$$

As usual, the listener's updated beliefs about the type of situation they are in, *t*, are a combination of their prior beliefs about the type of situation $p(t)$, and how well their prior experience with each type predicts the observations *x*, $p(x \mid t)$. In this formalization, observations and types are assumed to be only indirectly related, via the generative model parameters θ : each type predicts a range of generative models, $p(\theta \mid t)$, and each value of the generative model parameters predicts a range of observations, $p(x \mid \theta)$. In order to get the *marginal* likelihood of the signal *x* under situation *t*, $p(x \mid t)$, the particular values of the generative model parameters have to be averaged out:

$$p(x|t) = \int p(x|\theta)p(\theta|t)\mathrm{d}\theta \quad (23)$$

The jargony term for this averaging is *marginalization*. And a natural consequence of marginalization is that it leads to a preference for the most specific prior experience which provides a reasonably good fit to the observed data. We can think of marginalization as essentially taking a weighted average of the likelihood of the observation x, $p(x \mid \theta)$, weighted by the type-specific probability of the generative model parameters $p(\theta \mid t)$. On the one hand, for a type of prior experience that leads to very specific beliefs—like a particular familiar talker—the predicted observations $p(x \mid \theta)$ are very similar over the range of generative models that are given reasonably high probability, and so will all either fit the data reasonably well or not. On the other hand, a type of prior experience that leads to less specific beliefs about generative models—like experience with all speakers of the language—predicts a lot of different generative models. Some of these will probably provide a good account of the observation x—high likelihood $p(x \mid \theta)$—but the vast majority will not, resulting in a low average (maringal) likelihood $p(x \mid t)$.

This trade-off between specificity and accuracy goes both ways. A vague (e.g. language-wide) prior over generative models will allocate *some* likelihood to nearly *any* observation, whereas a specific (e.g. familiar talker) prior predicts only a very limited range of observations, and so assigns very high marginal likelihood to things in that range and basically nothing outside of it. So for observations that fall outside the range predicted by previous experience with specific talkers, less specific, group-level experience provide the best marginal likelihood, even if it's not (in absolute terms) very high.

Of course, inferring the type of prior experience that is relevant is only the first step to robust speech perception. Additionally, the listener also needs to infer the generative model that is best suited to the current situation. We can think of this as inference of the *joint posterior* distribution of generative model parameters and type of situation, $p(\theta, t \mid x)$. This combines inference about the type of situation—as discussed in this section—with situation-specific belief-updating—discussed above:

$$p(\theta, t|x) = \underbrace{p(\theta|x,t)}_{\text{situation-specific updated beliefs}} \underbrace{p(t|x)}_{\text{type of situation}} \quad (24)$$

Alternatively, we can think of just the overall updated beliefs about the generative model $p(\theta \mid x)$, which—like the prior beliefs $p(\theta)$—are a mixture of updated situation-specific beliefs, weighted by how likely each situation is given the observation:

$$p(\theta|x) = \sum_t p(\theta, t|x) = \sum_t p(\theta|x,t)p(t|x) \quad (25)$$

This brings us back to the central point of the second part of this paper: the ideal adapter framework reveals that we can look at recognition (and the resulting talker-specificity), adaptation, and generalization in speech perception as the natural consequence of a system that is trying to do the best that it can to comprehend speech in a variable—but structured—world. Because of the variability (or subjective non-stationarity) of the statistics of the speech signal, the listener has to adjust the generative model they use. There are two sources of information that listeners have when trying to do this: the speech they are currently

processing, and their previous experience and prior expectations. Different strategies arise from the fact that in each situation, the prior experience that is relevant—if any—is sometimes more and sometimes less informative about the appropriate generative model than the current speech itself. When the listener has no relevant prior experience, the best they can do is to try to learn the generative model by adapting their beliefs to better match the recently observed speech statistics, leading to rapid adaptation. When they have prior experience that is *directly* relevant, like with a familiar individual talker, the best thing they can do is to rely on that prior experience, removing the need to adapt and leading to talker-specific speech perception. When they have experience that is relevant, but not directly, then the best they can do is apply that experience to the current situation and then use the actual speech statistics to converge on an appropriate generative model.

For types of experience that correspond to very specific beliefs—like a highly familiar talker—the situation-specific beliefs will change little with the addition of one more observation. If such a component for a familiar talker provides the best match to the observed speech, then the result of this process of belief updating with a clustered mixture prior is that the listener has recognized and deployed their talker-specific generative model. If, on the other hand, the best-fitting type is a very broad, general type like all speakers of English, then the updated beliefs are dominated much more by the actual observed data then the prior, just like when the listener knows ahead of time that they need to adapt rapidly. Similarly, if prior beliefs of intermediate specificity dominate, then the updated beliefs will reflect the observed data more than they would for a highly familiar talker, but the prior beliefs will contribute more than in the most flexible case. This is how the ideal adapter is able to flexibly deal with the lack of invariance, making the best use of different types of prior experience with different generative models. By learning about and representing the *distribution* of generative models in the world, the ideal adapter is better prepared to infer the appropriate generative model across many different types of situations.

Open questions—We have argued here that the ideal adapter predicts that listeners dynamically balance flexibility and stability by *inferring* what kind of prior experience is currently most relevant. If the listener decides that experience with a particular familiar talker is most relevant, they will show stable, talker-specific speech perception. If, on the other hand, they decide that the only relevant prior experience they have is with the full distribution of talkers of the language, then they will rapidly and flexibly adapt.

The biggest open question that this raises is how—and even whether—listeners combine top-down and bottom-up information about what kind of prior experience is relevant. There are, in our view, at least two sides to this question. The first side is what a truly ideal adapter would do. The ideal adapter framework as currently posed addresses this only in the most general sense, saying that listeners *should* combine top-down and bottom-up information to determine what prior beliefs to use. But much more work is required in order to make further testable predictions. The second side is empirical, and in the rest of this section we review two empirical questions we see as particularly pressing.

What role do expectations of generalizability play?: The studies by Kraljic and Samuel (2005, 2006, 2007) suggest that listeners group together talkers when the distributions of

cues they produce are similar. But in all these studies, the similarity of the cue distributions is confounded with the overall similarity between male and female talkers in general on the relevant phonetic dimensions: male and female talkers don't systematically differ in their VOT distributions in American English (Allen et al., 2003), whereas they do differ in their frication frequency distributions (Newman et al., 2001). Thus listeners may have already made up their minds about whether to group the two talkers together before hearing anything. There's some evidence that top-down expectations and bottom-up similarity can be dissociated (C. M. Munson, 2011; Reinisch & Holt, 2014), but it remains to be seen exactly how listeners balance these two sources of information, both in the short term (e.g. over the course of a single experimental session) and over the long term (over days, weeks, or longer).

When are two generative models similar enough to group?: Our discussion of how listeners might generalize experience from one talker to another on the basis of similarity of their generative models glosses over a very basic point: even if their VOT ranges are overall very similar, the whole generative model for a male talker is *very* different from that of a female talker. Thus, while it might make sense to lump together experience with a male and female talker for the purposes of inferring their VOT distributions, it definitely does not make sense to lump together beliefs about their vowels. One way that this dilemma could be resolved is if listeners have extracted some information about the variance of different parameters in groups that contain males and females (like the ad hoc grouping that might be inferred during an experiment). Properties that vary a lot across talkers within a group are not very informative about particular talkers, whereas properties that are consistent across talkers in a group can be used to constrain beliefs about individuals much more strongly. Because males and females—within most groups—vary a lot in their vowels, but not so much in their VOT, listeners can use information about the VOT distributions of individual talkers to (on average) reliably infer properties of the group, which in turn are (on average) highly informative about properties of other talkers in the same group.

An alternate possibility is that listeners have more or less separate beliefs about different "chunks" of the generative model, potentially grouping talkers differently when considering beliefs about their VOT distributions than they do when considering beliefs about vowel formants. It is not immediately obvious what—if any—divergent predictions these two accounts would make, and more work is required to understand both their computational implications and related human behaviors.

Part II conclusion

We have argued that a range of strategies that human listeners use to deal with the lack of invariance all correspond to the best that they can do in a world where generative models vary from one situation to the next, but do so in a structured way that is tied to indexical variables like talker identity and group membership. The ideal adapter framework formalizes this notion by representing situation-specific beliefs about the range (distribution) of generative models that are expected in different types of situations, from very specific (an individual talker) to very general (an entire language).Although it is beyond the scope of this paper to discuss in any detail, this perspective can be applied to second (and third, etc.)

language acquisition as well (Pajak, Fine, Kleinschmidt, and Jaeger, *under review*). This range allows the listener to balance stability and flexibility, relying on prior experience when it is available and relevant, falling back on less informative but more flexible group-level beliefs. This system is not fail-safe. In order to achieve robust recognition of familiar talkers and guiding generalization across talkers, the speech perception system is forced to rely on prior experience. In highly atypical situations –which, fortunately, are bound to be rare– this very same property can slow down or prevent successful adaptation and perception (as also evident in second language learning).

Explicit structured beliefs are one way of achieving sensitivity to structured variability of cue-category mappings. However, the primary take-home message from Part II is that sensitivity to this structured variability is a critical feature of any model that seeks to explain the robustness of human speech perception. Formulating such a model and evaluating its predictions remains a task for future work.

Part III

We close our paper with a discussion of how the present work relates to other approaches and issues. We chose to take a *computational-level* approach to understanding how the speech perception system deals with the lack of invariance. Such an analysis focuses on the *problem* of robust speech perception. That is, our approach is guided by questions about the 'why', or the purpose of the system, the goals it typically serves, and the world it has to function in. Focusing on the 'why' of speech perception does not directly address the actual cognitive processes (and neural mechanisms) that must carry it out (Marr, 1982), but it has allowed the development of models that provide a good fit to human behavior, both in speech perception (Clayards et al., 2008; Feldman et al., 2009; Norris & McQueen, 2008; Sonderegger & Yu, 2010) and in a range of phonetic adaptation phenomena (Part I of this paper). It also has allowed us to make novel predictions, some of which we have tested here and the rest of which we hope will guide future research and the development of cognitive and neural models. The ideal adapter framework is *normative*, in that it looks only at the in-principle constraints on performance that come from the inherent difficulty of the task and the limited information that is available in the world, without considering resource limitations. The brain is obviously not unbounded in its resources, and considerations that come from processes, representations, and mechanisms are thus relevant to future development of the framework presented here.

However, this gulf is not as wide as it might first appear. Neurally plausible algorithms exist for the sort of inferences required by the ideal adapter framework (Beck, Pouget, & Heller, 2012; Friston, 2005; Rao & Ballard, 1999), and resource limitations can be included in a normative framework. For instance, recent work has focused on developing "rational approximations" to rational (normative) models like the ideal adapter. These models ask what the best possible performance is under different types and levels of cognitive or neural resource constraints (Griffiths, Vul, & Sanborn, 2012; Sanborn et al., 2010; Shi et al., 2010). For instance, one notable constraint on normative, Bayesian inference is the amount of uncertainty—or the number of different hypotheses—that can be simultaneously maintained. Listeners can apparently maintain enough uncertainty about whether or not two talkers

should be grouped together to overcome initial biases about how to generalize phonetic recalibration (Kraljic & Samuel, 2005, 2006; C. M. Munson, 2011). However, listeners also seem to prefer to stick to "first impressions" more than a purely normative model would predict (Kraljic et al., 2008). This suggests limits on how much uncertainty listeners are willing or able to maintain (see also Tzeng, Alexander, Sidaras, & Nygaard, 2014 for related effects of presentation order when generalizing across talkers).

Likewise, process- and mechanistic-level theories do not exist in a vacuum, and computational-level considerations about the task they must perform can be relevant and informative. In this final part of the paper, we will expand on some particular ways that we think the ideal adapter framework relates to and potentially informs a range of other approaches and topics. First, a number of other approaches to the lack of invariance have been proposed, and we will briefly lay out where we think the ideal adapter framework differs from, conflicts with, and complementes these other approaches. Next, we turn to how the proposed framework, and the body of empirical literature it unites, might inform the debate about the underlying representations used by speech recognition, and the role of subphonetic detail in phonetic representations. Finally, we will discuss the possible implications for this work beyond speech perception by adult listeners to both higher-level language processing and the acquisition of phonetic categories during development.

How does the ideal adapter relate to other approaches to the lack of invariance?

Our approach is novel when compared to most models of speech perception—Bayesian and otherwise—that do not learn at all (Clayards et al., 2008; Feldman et al., 2009; McClelland & Elman, 1986; Norris, 1994; Norris et al., 2000). But we are hardly the first to consider the problem of the lack of invariance in speech perception, and a variety of other approaches exist that either directly address flexibility in phonetic categorization or address related problems. We review these in this section.

Existing models of speech perception and adaptation—The existing models that are most directly related to our approach deal with plasticity in phonetic categorization by adding slower learning (and habituation) dynamics to existing connectionist or dynamical systems models of phonetic categorization (Lancia & Winter, 2013; Mirman et al., 2006). In both of these models, learning is modeled with a Hebbian learning mechanism that increases the feedforward connection weight between acoustic/phonetic input units to categorical output units based on repeated co-activation. Such a learning mechanism is sufficient to qualitatively capture phonetic recalibration, but it is not clear whether the generally slow temporal dynamics of Hebbian learning can account for the very rapid recalibration effects that are typically observed (Guediche, Blumstein, Fiez, & Holt, 2014). Furthermore, without an additional habituation mechanism (which *is* present in Lancia & Winter, 2013) Hebbian learning may not be sufficient to capture selective adaptation.

At a deeper level, existing models of plasticity in speech perception are—in principle or in practice—generally "flat" learners, which (in the language of the ideal adapter framework) adapt a single set of beliefs about the generative model based on recent experience. This includes connectionist models of phonetic adaptation (Lancia & Winter, 2013; Mirman et

al., 2006) as well as distributional learning models of the *acquisition* of phonetic categories (e.g. Feldman, Griffiths, et al., 2013; McMurray et al., 2009; Vallabha et al., 2007). Because none of these models represent the fact that the distributions (or connection weights from input to output units in Lancia & Winter, 2013; Mirman et al., 2006) are not only situation-specific but might be encountered again after an interruption, these models cannot account for the persistence of adaptation over intervening exposure (as in Eisner & McQueen, 2006; Kraljic & Samuel, 2005). These models also cannot explain when we do or do not generalize across talkers (Kraljic & Samuel, 2005, 2006, 2007).[30] The belief-updating model presented in Part I falls into the same class of flat learner models. Thus, one of critical goals for future computational work on speech perception is to continue the development and test of the full ideal adapter framework, as we have begun to outline in Part II.

The remaining class of approaches to plasticity in speech perception is the class of episodic or exemplar models (e.g., Goldinger, 1998; Johnson, 1997b, 2006; Pierrehumbert, 2003). These models are arguably motivated by many of the same considerations that motivate the ideal adapter framework: the mapping between cues and categories is variable, in often idiosyncratic ways (Johnson, 2006), and so the speech perception system has to be sensitive to this variability. Exemplar models achieve this by storing raw acoustic traces for each exemplar that has been encountered, and categorize new inputs based on similarity to stored exemplars. By remembering every instance of a category, these models (implicitly) learn the corresponding distribution of sounds, and can achieve persistent talker-specific representations (Goldinger, 1998; Johnson, 2006). However, simply storing raw episodes alone is not sufficient to explain the ways that human listeners generalize learning on one phonetic category to unheard words (Cutler, Eisner, McQueen, & Norris, 2010; McQueen et al., 2006), other contrasts (Kraljic & Samuel, 2006), and other talkers (Kraljic & Samuel, 2006, 2007; Reinisch & Holt, 2014). For such flexibility, some additional sensitivity to the underlying structure of the variation in cue mappings is required, both at the level of how the acoustic signal is analyzed into linguistic units (Cutler et al., 2010) and at the level of how different talkers can be grouped based on their generative models (Part II of this paper). To some extent this, too, has been recognized in recent work on episodic theories of language processing (e.g., Johnson, 2006, 2013; Pierrehumbert, 2003; van den Bosch & Daelemans, 2013).

To illustrate this point further, we have argued (in Part II) that the speech perception system can benefit from structure in the world in how generative models vary across situations. This structure means that prior experience with the same or similar generative models can inform future adaptation and processing. We have discussed how, in the ideal adapter framework, sensitivity to this structure can be naturally formalized by structured *representations*, in the form of talker-and group-specific beliefs about (distributions over) generative models.

Memory-based or episodic theories provide an alternative approach which is superficially conflicting. In these approaches, prior experience is *represented* in an unstructured way, but

[30]Mirman et al. (2006) accounts for the results of Kraljic and Samuel (2005, 2006) by using cue representations that encode the similarity or dissimilarity of the two talker's test continua. But as far as we can tell, this approach fails to predict the opposite results (C. M. Munson, 2011; Reinisch & Holt, 2014)

sensitivity to the structure of that experience might still arise through *processes* by which similarity to stored episodes is computed (e.g., analogical reasoning van den Bosch & Daelemans, 2013). This is an area of ongoing research, and it is an interesting question whether these two approaches—structure by representations or by processes—make substantively different predictions in principle. It is possible, for example, that analogical reasoning applied to unstructured memory traces can implement the sort of structure-sensitive computations we have argued for.

What counts as a situation for adaptation?—In our discussion of how generative models vary from one situation to the next, we have focused on talkers (or other indexical variables) as the main driver of this variability. The reason for this emphasis on talkers is that differences between talkers are responsible for a large amount of variability in how phonetic categories are realized acoustically (Allen et al., 2003; Hillenbrand et al., 1995; McMurray & Jongman, 2011; Newman et al., 2001, among others). This lead us to hypothesize that listeners' beliefs about the generative model of speech need to be thought of as conditioned on their beliefs about the talker (or type of talker) they think they are listening to.

However, talker differences are not the *only* source of the lack of invariance, and nonlinguistic factors like background noise or general acoustics can also change from situation to situation. Additionally, the realization of phonetic categories depends on the *linguistic* context that those sounds are produced in (Liberman et al., 1967), and a variety of models have been proposed to account for how categorization of one segment is affected by adjacent segments (Massaro, 1987; Nearey, 1997; Nearey & Assman, 1986; Oden & Massaro, 1978; Smits, 2001; Sonderegger & Yu, 2010).

This raises an important question: is what we have called a "situation" (understood as the non-linguistic aspects of the current situation like talker or setting) different from the notion of "context" in, for instance, models of coarticulation? The ideal adapter framework points towards a way of approaching this question. This is based on the underlying statistical properties of how the cue-category mappings represented by the generative model vary based on talker (and other indexical variables), linguistic context, and the combination of the two. If the effects of linguistic context are themselves sufficiently variable *across* talkers (as experienced by the listener), then the ideal adapter framework says that the speech perception system has a strong incentive to track category statistics as a joint function of *both* linguistic and non-linguistic context. If not, then they can be safely tracked separately. Even if the effects of linguistic context and indexical situation can be, computationally, treated as basically the same, it is a separate question of what the brain makes out of these two very different kinds of variability. The effects of linguistic context are relatively localized in time (only a couple of syllables), whereas indexical context is much more diffuse (an entire discourse, or even multiple separate encounters), and so each may impose very different demands on memory and other processing resources.

What is the nature of representations underlying speech recognition?

Even though our work here is pitched at a computational level, it does have implications for existing linguistic theories of speech recognition. By relating the lack of invariance to the strategies that people use to achieve robust speech perception—adaptation, talker-specificity, and generalization—the ideal adapter framework highlights shortcomings in the starting points for existing theories of how the speech recognition system maps cues onto linguistic categories. We will argue that the ideal adapter framework also helps to understand how recent moves away from these more extreme starting points are a step in the right direction, and—we hope—points to further productive directions for future research.

Abstractionist and episodic approaches to speech recognition—One of the persistently debated issues in speech recognition is the degree of abstraction in the representations which mediate cue-category mappings. At one extreme, *abstractionist* models posit that speech perception is mediated by prelexical representations that strip away subphonetic variation, discarding acoustic information which is not relevant for making (possibly probabilistic) categorical distinctions (McClelland & Elman, 1986; Norris, 1994; Norris et al., 2000; Norris & McQueen, 2008). At the other extreme, *episodic* models posit that speech perception is a direct mapping of detailed acoustic traces to linguistic units (often lexical items), preserving all of the fine-grained acoustic information (Goldinger, 1998; Johnson, 1997b, 2006; Pierrehumbert, 2002). These two perspectives constitute dramatically different approaches to the lack of invariance, each of which is insightful but also falls short in different ways.

On the one hand, in the face of the lack of invariance abstractionist theories have historically relied on an explicit process of “talker normalization” where acoustic cues are normalized by pre-linguistic perceptual processes such that a single set of cue values or template can be used for each category, regardless of the talker. For instance, vowel formant frequencies, which vary systematically with talker gender (Hillenbrand et al., 1995; Peterson, 1952) might be normalized with respect to the values of other formants, or the fundamental frequency (Strange, 1989). More recent approaches include what we termed “flat learner” models above, where a single set of cue-category mappings is dynamically adjusted based on recent input via something like Hebbian associative learning (Lancia & Winter, 2013; Mirman et al., 2006).

On the other hand, episodic theories avoid the problem of talker normalization altogether, because the detailed acoustic traces of, for instance, individual word or vowel tokens are stored separately, and recognition proceeds on the basis of overall acoustic similarity. By storing large enough (e.g. word-sized) acoustic traces, each remembered episode encodes enough information to be useful in distinguishing individual talkers, supporting talker-specific cue-category mapping (Goldinger, 1998; Johnson, 1997b; Pierrehumbert, 2002). For unfamiliar talkers, new episodes are recognized based on some measure of overall similarity to stored episodes from other talkers, allowing generalization across talkers.

Insights and problems of each approach—We can think of these as two extreme endpoints for how to map from acoustic cues to linguistic representations. Abstractionist

accounts seek a *minimal*(single) set of cue-category mappings (possibly to be tuned and re-tuned based on experience), while episodic accounts use a *maximal* set of cue-category mappings, one for every observation ever made. Each approach is founded on useful insights. Abstractionist models are informationally efficient, collapsing all the variability across situations into a compact set of sublexical cue-category mappings which supports efficient generalization across lexical items. Episodic models are infinitely flexible, and can pick up on any informative set of cues simply by virtue of tracking all of them.

But each approach also comes at a cost. For abstractionist accounts, normalization is hard, if not impossible, given that much of the variability across talkers is not due to fixed (e.g. physiological) factors but rather stylistic and thus needs to be learned (Johnson, 2006; Pardo & Remez, 2006). The Hebbian learning mechanisms that have been proposed typically operate at much longer time scales than the very rapid recalibration that is commonly observed (Guediche et al., 2014; Mirman et al., 2006). Even if the learning rate can be dynamically adjusted, existing proposals for "flat" learners cannot account for the speech perception system's sensitivity to structure (e.g. persistent talker-specific representations). Episodic accounts which track raw acoustic episodes of a large enough size to simultaneously encode the phonetic features and the other acoustic information that tells you how to interpret those features (e.g. frication energy and adjacent vowel formants/f0 to get gender) cannot easily generalize recalibration to unheard words (Cutler et al., 2010). More importantly, episodic models cannot generalize across groups of different levels of specificity by simply recording acoustic episodes. Rather, they require some additional mechanism for "tagging" and filtering or weighting exemplars based on indexical variables.[31]

Charting a middle course—Of course, these criticisms are based on extreme or purist interpretations of these two approaches, and there is no reason why abstractionist models cannot be made to be a bit more like episodic models and vice-versa. Indeed, this has been the trend in recent years, and a number of hybrid models have been called for or proposed (e.g., Ernestus, 2014; Goldinger, 2007; McLennan, Luce, & Charles-Luce, 2003). Of particular interest are proposals for abstractionist models that learn (Lancia & Winter, 2013; Mirman et al., 2006), and episodic models that track episodes at sublexical granularity (Pierrehumbert, 2006) or tag episodes with explicit indexical variables and use these to bias later recognition (Johnson, 2006, 2013).

In fact, the perspective offered by the ideal adapter suggests that some balance between complete abstraction and complete lack of abstraction is optimal. The particular balance depends both on the *current situation*(which might require more or less flexibility and hence sensitivity to individual episodes) and on the listener's previous experience. A listener who has experienced input from a broad range of different generative models will have reason to believe that flexibility will be required in the future, and thus will be more likely to display it. A Listener who has experienced a narrower range or more consistent groupings has required less flexibility. The ideal. adapter framework relates an individual listener's

[31]Or, alternatively, a "smarter" similarity function that is somehow sensitive to the structure of the episodes, both linguistically and indexically (e.g., as discussed above, van den Bosch & Daelemans, 2013)

(potentially idiosyncratic) structured experience with different cue-category mappings to the kind of representations that will best serve the listener's needs, interpolating between complete abstraction and no abstraction. Like the rational model of categorization (Anderson, 1991), the ideal adapter framework proposes that listeners tune the specificity of their cue-category mappings based on their beliefs about the underlying structure of the world.

The ideal adapter framework thus incorporates insights from both abstractionist and episodic approaches. On the one hand, like episodic approaches, it recognizes the need to estimate cue distributions. On the other hand, it recognizes that there is a substantial benefit from having only a compact set of generative model parameters that need to be adapted in each situation. In this sense, our approach is inspired by normalization accounts. However, instead of adjusting the *acoustic cue space* to make all talkers align with a fixed set of cue-to-category mappings, the ideal adapter adjusts the *category-to-cue mappings* for each situation. We have also tentatively proposed that there is a parallel trade-off at a higher level, in estimating the distribution of generative models themselves across situations. This suggests that listeners might be able to use exemplars in generative model parameter space (rather than acoustic space) to estimate, in an approximate, boundedly-rational way, the overall distribution of generative models across situations (Ashby & Alfonso-Reese, 1995; Gibson et al., 2013; Griffiths, Sanborn, Canini, & Navarro, 2008) and to adapt to the current situation (Shi et al., 2010). However, there is also a benefit from abstracting away from individual situations by explicitly representing summaries of different group-level distributions in order to generalize across groups. While there is some evidence that phonetic adaptation reflects updating of top-down category-to-cue representations rather than bottom up cue warping (Dahan, Drucker, & Scarborough, 2008), further work is required to determine whether this is what human listeners always do and to computationally flesh out these different approaches.

The role of subphonetic detail—The importance of subphonetic detail for *categorizing* speech has long been acknowledged, both by probabilistic, ideal observer models (Clayards et al., 2008; Feldman et al., 2009; Norris & McQueen, 2008; Sonderegger & Yu, 2010) and others (McClelland & Elman, 1986; Norris, 1994; Norris et al., 2000). For example, subphonetic detail is necessary for making linguistic inference in the face of variability and uncertainty. Higher levels of analysis benefit from knowing not just which phonetic category is *most* likely, but how certain that judgement is and what plausible alternatives are. To the extent that uncertainty can be carried through the different levels of inference, later evidence can be integrated rationally, allowing the listener to revise earlier interpretations when necessary. So-called right-context effects in spoken language processing show that listeners do, in fact, maintain uncertainty over the categorization of previous input (Bard et al., 1988; Connine et al., 1991; Dahan, 2010; Grosjean, 1985), and there is some evidence that the same occurs in written language processing, too (Levy, Bicknell, Slattery, & Rayner, 2009). Conversely, people are sensitive to information from anticipatory coarticulation (Dahan, Magnuson, Tanenhaus, & Hogan, 2001; Marslen-Wilson & Warren, 1994; McQueen, Norris, & Cutler, 1999; Whalen, 1984), where upcoming segments change how an earlier

segment is produced, providing probabilistic information about how to categorize *later* input and the underlying words.

The ideal adapter makes an even stronger claim about the importance of subphonetic detail. Compared to just speech recognition, a different (but overlapping) type of subphonetic detail is necessary in order to achieve robust categorization in the face of situational variability. We have argued that the distributions of cues corresponding to a category change across situations, and that robust categorization requires that these distributions be tracked. Such distributional learning relies on predictions about the subphonetic distribution of cues expected based on prior experience in order to detect when, and determine how, the current distributions deviate from what is expected. Two subphonetic variants of a particular category which would be classified—even probabilistically—in exactly the same way can nevertheless in principle produce very different *adaptation* effects because they signal different underlying distributions. That is, in order to effectively update beliefs about the distribution of cues for each category, the speech perception system needs to be sensitive, at some level, to types of subphonetic differences that are irrelevant for classification.

Moreover, because different talkers can produce different distributions (which are distinguished by potentially subtle subphonetic differences), subphonetic detail provides information about *who* is talking (Creel & Bregman, 2011; Pardo & Remez, 2006). In a world where phonetic category distributions vary systematically as a function of who is talking, any information about the talker, even at a general level like gender or language background, is informative about what kind of category distributions will most likely be appropriate. This not only determines how adaptation proceeds but also can determine how speech sounds ought to be categorized. Much of this information is contained in subphonetic details like voice quality (Pardo & Remez, 2006), but also, as we have argued, in subphonetic variation in phonetic cues themselves. Thus, our analysis suggests that even at fairly high levels of processing, the speech recognition system should be sensitive to subphonetic variation even when this variation is not directly informative about categorization.

Implications of the ideal adapter framework beyond speech perception

We now turn to the broader implications of this framework for speech perception, acquisition, and language comprehension. We begin by discussing the relation between adaptation in adults and language acquisition in infants. We then discuss how the proposed framework highlights an important parallel between learning (in the form of adaptation) and processing. Following that, we summarize parallels that the proposed account highlights between speech perception and language processing beyond speech perception. Finally, we suggest how our computational analysis of speech perception and the resulting models might be realized as process and mechanistic models, and some of the challenges in doing so.

Parallels between acquisition and adaptation: Both can be understood as a form of distributional learning—We have proposed that adaptation is a form of distributional learning and hence that distributional learning is a central part of *processing* spoken language by adult listeners in the face of variability across situations. Distributional

learning has also been proposed as a mechanism by which linguistic categories are learned during language acquisition in infants (e.g., Aslin, Saffran, & Newport, 1998; Gómez & Gerken, 2000; Saffran, Aslin, & Newport, 1996; Wonnacott, Newport, & Tanenhaus, 2008) and adults (e.g., Pajak & Levy, 2011; Saffran, Newport, & Aslin, 1996). This includes the acquisition of phonetic categories (e.g., McMurray et al., 2009; Toscano & McMurray, 2010; Vallabha et al., 2007). For example, infants have demonstrated sensitivity to the distribution of acoustic cues to phonetic categories as early as 6 months of age (Maye et al., 2002). This parallel has been taken to suggest that language acquisition and adaptation can be attributed to the same implicit learning mechanisms, which continue to operate throughout life (e.g., Botvinick & Plaut, 2004; Chang et al., 2006; Elman, 1990). We briefly review the arguments and challenges for such accounts.

While the adult listener has uncertainty about the distribution of cues associated with each category, the language learner faces the additional challenge of not knowing what the underlying categorical structure is. On top of the distributions (e.g. mean and variance), the language learner must therefore infer the existence and number of categories. Additionally, infant language learners must arguably do so from unlabeled data, so that at least the earliest stages of language acquisition require *un* supervised learning. Given these *prima facie* differences between adaptation and acquisition, it is striking that distributional learning models similar to the one we propose here have been found to account for phonetic category acquisition data (Feldman, Griffiths, et al., 2013; McMurray et al., 2009; Vallabha et al., 2007), as well as acquisition of other linguistic structures (M. C. Frank, Goodman, & Tenenbaum, 2009; O'Donnell, Snedeker, Tenenbaum, & Goodman, 2011).

This and the work we present here have two implications for future work. First, our analysis of adaptation as central to speech processing suggests that the distributional learning mechanisms that seem to underly acquisition might continue to operate throughout life (as also proposed in Botvinick & Plaut, 2004; Chang et al., 2006; Elman, 1990). Adult listeners have acquired rich distributional knowledge about both the cue-to-category mappings (generative model) of their native language and the distribution of these mappings across talkers. Still, even adult listeners are routinely exposed to situations in which they need to learn novel statistics. In ongoing work, we have found that a model which correctly infers that there are two voicing categories based on unlabeled distribution of VOTs can *also* account for adult listeners shifts in categorization after hearing unlabeled, shifted distributions of VOTs, and strikingly can do so using a single set of parameters (Toscano, Munson, Kleinschmidt, and Jaeger, *under revision*). Thus it appears that *some* distributional learning may be involved from acquisition to adult language use, echoing life-long learning accounts.

In this context, it is worth noting that research on phonetic category acquisition has generally employed unsupervised learning models. The model we used in Part I was supervised. This is, however, not an in-principle limitation of the ideal adapter framework. Indeed, adults clearly can engage in unsupervised adaptation (Clayards et al., 2008; C. M. Munson, 2011). Further work is required to better understand the relative role of supervised and unsupervised learning (or the continuum between these extremes; e.g., Gibson et al.,

2013; Zhu, Rogers, Qian, & Kalish, 2007) both during acquisition and in speech perception in adults.

Second, as we have pointed out, the distributional statistics of phonetic categories depend on talkers, accents, etc. (e.g., Figure 1). This raises the question of how infants acquire categories statistically unless they separate the observations they make based on who is talking. This is an underexplored question that deserves further attention in future work. Preliminary evidence comes from White and Aslin (2011), who find that 20-month-old toddlers demonstrate phonetic recalibration in paradigms similar to those used in research on adults. This suggests that infants may be learning talker- or situation-specific phonetic category representations, as we have argued for adults. Furthermore, some computational studies demonstrating the ability of distributional learning algorithms to extract categorical structure have used input data that comes from multiple talkers, even collapsing across gender (e.g., Feldman, Griffiths, et al., 2013). This increases category overlap, which *a priori* would seem to make it more difficult to learn good phonetic categories. If infants separate tokens according to who produced them (or even according to non-phonetic features like voice quality) then at least some of this difficulty may be overstated.

However, tracking talker-specific distributions of cues also means that learners need to track a set of speech sound statistics for each individual talker, instead of just a single set of language-level distributions (as is typically modeled). This introduces two types of costs. First, it requires more statistics to be represented (remembered). Second, there are fewer observations available for each (talker-specific) generative model that needs to be learned, decreasing the reliability of these models. One way to ameliorate both costs is to recognize generalizations across talkers. In any case, however, the costs introduced by talker-specific or group-specific expectations can be justified when the variation *between* talkers is so large that the differences in the cue distributions of phonetic categories *within* each talker are obscured. Future work could employ computational simulations to probe the statistical conditions under which this trade-off is predicted to work out in favor for talker- and group-specific learning. Similarly, more experimental research is required on the changes in the way we learn and generalize throughout language acquisition.

Parallels between processing and learning: Both involve inference—Our ideal adapter framework is inspired by previous work which treats speech perception as a problem of inference under uncertainty (Clayards et al., 2008; Feldman et al., 2009; Norris & McQueen, 2008; Sonderegger & Yu, 2010). Like this previous work, we analyze the problem of speech perception via an ideal listener model where processing depends on the listener's model of the statistics of phonetic categories. However, the models developed in this previous work generally assume (tacitly) that for the purposes of adult language processing, these statistics are known and fixed. Our analysis points out that this is not a reasonable assumption, because the lack of invariance introduces variability in these statistics across situations.

In the proposed framework, the lack of invariance means that learning is closely intertwined with processing of spoken language. One insight of the ideal adapter is that processing and learning are both instances of inference under uncertainty. An ideal listener uses knowledge

of how phonetic categories generate acoustic cues in order to determine which underlying category best explains an observed cue. In the same way, an ideal adapter uses knowledge of how different generative models produce different distributions of acoustic cues in order to update their beliefs about the appropriate generative model for the current situation. That is, in the ideal adapter framework, adaptation is inference at another level, inference of generative models.

There's a deeper parallel between processing and learning in speech perception, as well. Phonetic categories don't occur arbitrarily, in isolation, but have a regular structure determined by, for instance, the words of the language. An ideal listener uses knowledge about this structure to narrow down the possible explanations for a given acoustic cue through the prior $p(c)$ (Feldman, Myers, White, Griffiths, & Morgan, 2013). In the same way, we have proposed that an ideal adapter should exploit the fact that changes in phonetic category distributions are systematically related to changes in talker, dialect, etc. This allows an ideal adapter to quickly and efficiently infer (re-learn) the distributions of a familiar talker when encountering them again, or to generalize from experience with talkers from a particular foreign accent group to a novel talker from the same group.

Language understanding beyond speech perception—The idea that language understanding involves prediction and inference under uncertainty has also guided work on language processing beyond speech perception. For example, expectation-based theories of sentence processing hold that comprehenders use knowledge of the statistics of syntactic structures in order to generate expectations about upcoming material (Hale, 2001; Levy, 2008a; MacDonald, Pearlmutter, & Seidenberg, 1994; MacDonald, 2013; Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995). As for speech perception, reliance on such statistics is efficient (Levy, 2008a; Smith & Levy, 2013). As predicted by these theories, processing of structures which are highly likely given, say, a particular verb is faster than an alternative structure which is less likely (Garnsey, Pearlmutter, Myers, & Lotocky, 1997; MacDonald et al., 1994; Staub & Clifton, 2006; Trueswell & Tanenhaus, 1994). Similarly, word-by-word processing times in reading have been found to increase with decreasing contextual predictability of the word (specifically, its contextual surprisal, Demberg & Keller, 2008; Hale, 2001; S. L. Frank & Bod, 2011)

As is the case for phonetic categories, these statistics vary across situations (see Fine et al., 2013, for references). Thus, the same argument that we have made here for phonetic categorization—that effective comprehension relies on good estimates of the talker's generative model and deviations from expected statistics will lead to changes in comprehension—applies to syntactic processing, as well. Indeed, recent studies show that comprehenders *do* adapt to changes in the statistics of syntactic structures, with repeated exposure to a previously rare structure facilitating processing of that structure (Fine, Qian, Jaeger, & Jacobs, 2010; Fine et al., 2013; Jaeger & Snider, 2013; Kaschak & Glenberg, 2004).

These syntactic adaptation effects have much in common with phonetic adaptation. They are both rapid, occurring after only tens of exposures to the critical structure (Fine et al., 2013; Kaschak & Glenberg, 2004), and there is some evidence that they are persistent, lasting over

multiple days (Wells, Christiansen, Race, Acheson, & MacDonald, 2009), and that unusual syntactic preferences can be explained away (Hanulíková, van Alphen, van Goch, & Weber, 2012). Consistent with the hypothesis that such adaptation generally serves to make language comprehension robust to systematic situational variability, Kamide (2012) found that comprehenders tracked talker-specific syntactic preferences (high-vs. low-attachment), and that this adaptation was reflected in online processing.

Further reinforcing the tentative parallels between syntactic and phonetic adaptation, belief updating models similar to the one presented here can quantitatively account for these syntactic adaptation effects (Fine et al., 2010; Kleinschmidt, Fine, & Jaeger, 2012). Intriguingly, as we found here, these models also fit best with low effective prior sample sizes.

Similar findings are emerging for prosodic processing (Kurumada, Brown, Bibyk, Pontillo, & Tanenhaus, 2013), phonotactic constraints (Chambers, Onishi, & Fisher, 2010; Dell, Reed, Adams, & Meyer, 2000; Goldrick, 2004; Warker & Dell, 2006), pragmatic processing (Grodner & Sedivy, 2011), and semantic interpretation, such as quantifier processing (Yildirim, Degen, Tanenhaus, & Jaeger, 2013). This suggests that not only do people use their previous experience with the statistics of their language to predict upcoming material in order to facilitate comprehension, but that they are tracking situation-specific statistics and tuning their expectations to reflect changes in such statistics. The current work thus contributes to a growing literature that considers language comprehension to be intimately tied to adaptation and implicit learning (see also Dell & Chang, 2014; Chang et al., 2006; Fine et al., 2013; Jaeger, 2013; Jaeger & Snider, 2013; MacDonald, 2013).

Perception and learning beyond language—The problem of achieving robust speech perception in many situations in the face of the lack of invariance is not unique to speech or language. Generally, agents (both people and other animals) need to act effectively based on sensory cues, and the mapping from cues to the appropriate actions (or interpretations, more generally) can vary quite a bit. There is a *very* long literature on how agents manage to act appropriately in a variety of situations. Much of this comes down to *learning* new or adjusted cue-outcome associations, and as such the majority of the classical behavioral paradigms expose animals to highly novel cue-outcome mappings, in order to better assess and understand learning. However, a recurring theme in this work is that learners have to function in a *multi-context* world, where the cue-outcome mappings are not simply drawn from a single random distribution but rather depend systematically on factors of the underlying context.

Much of the work on such multi-context learning treats it as a change-detection problem, where the learner's strategy is to detect when the context has changed and then change their learning strategy by, for instance, increasing the learning rate (e.g., Courville, Daw, & Touretzky, 2006; Gallistel et al., 2001). The ideal adapter framework offer an alternative to change detection models (see also related proposals discussed in Qian et al., 2012). By trying to infer the underlying generative model (or cue-outcome mapping) in each situation, changes in context can be detected implicitly. Further, by tracking context-specific

generative models, previous experience can be used to efficiently re-learn previously encountered contexts.

There is indeed some evidence that learners do not abandon previously learned associations when learning new associations. For instance, extinction (unlearning) of a classically conditioned response (e.g. a fear response or eyeblink) is not absolute, and the original conditioned association is still present and can re-appear under a variety of circumstances from re-introduction of the original context, very brief exposure to the conditioned association, or simply spontaneously (Thanellou & Green, 2011; Bouton & King, 1983; Sissons & Miller, 2009). Similarly, early exposure to perturbed audio-visual spatial correspondences in barn owls leads to more rapid re-adaptation to the same perturbation later in life, even with substantial un-perturbed experience in between (Körding, Tenenbaum, & Shadmehr, 2007; Knudsen, 1998; Linkenhoker, von der Ohe, & Knudsen, 2005). Similarly, talker-and group-specific expectations in speech perception could be seen as resulting from context-sensitive associative learning, although we have not described it as such.

All of this raises the question as to whether the very same computational framework developed here—the ideal adapter—could be directly applied to the problem of domain-general learning in a multi-context world. It is not entirely clear that the entities of the ideal adapter can be mapped to more general settings, even if the central idea is applicable. What is, for instance, the analogue of a talker, or a phonetic category?

Speech perception as a computational problem is characterized by two, largely orthogonal, types of structure: linguistic and indexical. Linguistic structure refers to the fact that acoustic cues signal the presence of underlying phonetic categories, lexical items, syntactic structures, etc. Indexical structure refers to the fact that these cue-category mappings—which we have referred to as the generative model—vary across situations based (in large part) on *who* is talking, and other indexical variables like accent, dialect, gender, etc. The fact that the same set of generative model parameters (e.g. mean VOT of /p/) are relevant for all (or nearly all) talkers means that listeners have a lot to gain by combining experience with different generative models in different situations. But the same may not be true of other domains, where the underlying categorical structure of the stimuli itself varies across contexts.

This is not to say that the basic logic of the ideal adapter framework does not apply: people (and other agents) can benefit from tracking the structure of how cue statistics vary across situations to the extent that this variation *is* structured. By the same logic, a learner should *not* build a model of how sensory statistics vary across situations if that variability is unpredictable. In such a domain, the ideal adapter predicts that people would *not* show situation-specificity, but rather continuously adapt.[32] Similarly, in a domain where the variability across situations is informative about both the cue-category mappings and what categories are relevant, then there is little benefit to tracking the distribution of category

[32]This leaves aside the possibility of any higher-level inference about whether or not situation-specificity is appropriate that people may be doing.

statistics across situations. That is, if each category is only relevant in one particular situation, then by the logic of the ideal adapter framework, people might not generalize across situations.

This is also an argument for speech perception as a good test case for theories of multi-context learning. Variability in speech *is* highly structured—both linguistically and indexically—so that listeners pick up on and use both kinds of structure.

Conclusion

The proposed ideal adapter framework provides a potential solution to one of the oldest questions in research on speech perception: how do listeners overcome the computational problem caused by the lack of invariance of the speech signal? Recent proposals treat speech perception as a problem of inference under uncertainty about the phonetic category (or other linguistic unit) that a talker intended to produce (e.g., Clayards et al., 2008; Feldman et al., 2009; Norris & McQueen, 2008). The ideal adapter extends these models by treating speech perception as a problem of inference under uncertainty at *multiple* levels. Robust speech perception requires that listeners continuously draw inferences not only about *what* the talker is trying to communicate, but also about the cue-category mappings that the talker is using (i.e., the talker's generative model). Moreover, in order to determine what previous experience is relevant in making these inferences, and how relevant it is, these inferences in turn depend on higher-level inferences about *who* the talker is. This ranges from specific talker identity—supporting *recognition* of familiar talkers—to more general groups like gender, accent, or dialect—supporting generalization across similar talkers. The proposed multi-level inference solution can capture a variety of otherwise puzzling properties of speech adaptation and provides a guiding framework for future research on speech perception, adaptation, and generalization.

The challenges posed by variability are not unique to speech perception, but rather general to the problem of effective perception and action in a variable world. This problem has been explored in the context of motor control (e.g., Körding, Beierholm, et al., 2007) and reinforcement learning (e.g., Cho et al., 2002; Gallistel et al., 2001), where it is typically cast as a problem of detecting changes in the statistics of the local environment (change detection). Our proposal highlights the fact that in a world where substantial parts of cross-situation variability are not random, but rather structured, simply detecting changes is not enough. Rather, learners can benefit from inferring the underlying structure to cross-situation variation, in order to recognize familiar situations and generalize to similar situations. In speech perception, the major source of variation across situations is the *talker*, but the same logic can be applied to other domains (Qian et al., 2012; Qian, Jaeger, and Aslin, *submitted*). The ideal adapter highlights the potential of speech perception to serve as a model organism for understanding perception in a variable—but structured—world, and suggests that superficially unrelated phenomena from non-linguistic perceptual/motor domains might be informative about language processing and acquisition and vice-versa.

Acknowledgments

We would like to thank Jean Vroomen for graciously providing data for model fitting, as well as original stimuli files for replication, and Andrew Watts for technical help with our own experiments.

We also thank Dennis Norris, Naomi Feldman, and two anonymous reviewers whose comments on earlier drafts substantially improved and clarified this paper. We are particularly indebted to everyone who has provided feedback on and criticism of previous version of this work; these include (but are not limited to): Richard Aslin, Delphine Dahan, Gary Dell, Lori Holt, Robert Jacobs, Roger Levy, Bob McMurray, Rajeev Raizada, Arthur Samuel, Michael Tanenhaus, Joe Toscano, and members of the Aslin, HLP, and Jacobs labs at the University of Rochester, as well as participants and audiences at meetings where previous versions of this work were presented (especially the Workshop on Current Issues and Methods in Speaker Adaptation at The Ohio State University in 2013); any errors remaining are our own.

This work was partially funded by an NSF Graduate Research Fellowship to DFK and NIHCD R01 HD075797 as well as an Alfred P. Sloan Fellowship to TFJ. The views expressed here are those of the authors and not necessarily those of the funding agencies.

References

Ajmera J, McCowan I, Bourlard H. Robust speaker change detection. Signal Processing Letters, IEEE. 2004; 11(8):649–651.

Allen JS, Miller JL. Listener sensitivity to individual talker differences in voice-onset-time. The Journal of the Acoustical Society of America. 2004; 115(6):3171. [PubMed: 15237841]

Allen JS, Miller JL, DeSteno D. Individual talker differences in voice-onset-time. The Journal of the Acoustical Society of America. 2003; 113(1):544. [PubMed: 12558290]

Anderson, JR. The adaptive character of thought. Hillsdale, NJ: Lawrence Erlbaum Associates; 1990.

Anderson JR. The adaptive nature of human categorization. Psychological Review. 1991; 98(3):409–429.

Andruski JE, Blumstein SE, Burton M. The effect of subphonetic differences on lexical access. Cognition. 1994; 52(3):163–187. [PubMed: 7956004]

Ashby F, Alfonso-Reese LA. Categorization as Probability Density Estimation. Journal of Mathematical Psychology. 1995; 39(2):216–233.

Aslin RN, Saffran JR, Newport EL. Computation of Conditional Probability Statistics by 8-Month-Old Infants. Psychological Science. 1998; 9(4):321–324.

Babel, M.; Munson, B. Producing Socially Meaningful Linguistic Variation. In: Ferreira, V.; Goldrick, M.; Miozzo, M., editors. The oxford handbook of language production. Oxford: Oxford University Press; 2014. p. 308-328.

Baese-berk MM, Bradlow AR, Wright BA. Accent-independent adaptation to foreign accented speech. JASA Express Letters. 2013; 133(3):174–180.

Bard EG, Shillcock RC, Altmann GTM. The recognition of words after their acoustic offsets in spontaneous speech: Effects of subsequent context. Perception & Psychophysics. 1988; 44(5):395–408. [PubMed: 3226889]

Beck JM, Ma WJ, Pitkow X, Latham PE, Pouget A. Not noisy, just wrong: the role of suboptimal inference in behavioral variability. Neuron. 2012; 74(1):30–39. [PubMed: 22500627]

Beck, JM.; Pouget, A.; Heller, K. Complex Inference in Neural Circuits with Probabilistic Population Codes and Topic Models. In: Bartlett, P., editor. Advances in neural information processing systems. Vol. 25. Red Hook, NY: Curran Associates, Inc.; 2012. p. 3059-3067.

Bejjanki VR, Beck JM, Lu Z-L, Pouget A. Perceptual learning as improved probabilistic inference in early sensory areas. Nature Neuroscience. 2011; 14(5):642–648. [PubMed: 21460833]

Bejjanki VR, Clayards M, Knill DC, Aslin RN. Cue integration in categorical tasks: Insights from audio-visual speech perception. PLoS ONE. 2011; 6(5):e19812. [PubMed: 21637344]

Bertelson P, Vroomen J, de Gelder B. Visual recalibration of auditory speech identification: a McGurk aftereffect. Psychological Science. 2003; 14(6):592–597. [PubMed: 14629691]

Botvinick MM. Hierarchical reinforcement learning and decision making. Current opinion in neurobiology. 2012; 22(6):956–962. [PubMed: 22695048]

Botvinick MM, Plaut DC. Doing without schema hierarchies: a recurrent connectionist approach to normal and impaired routine sequential action. Psychological Review. 2004; 111(2):395–429. [PubMed: 15065915]

Bouton ME, King Da. Contextual control of the extinction of conditioned fear: tests for the associative value of the context. Journal of experimental psychology. Animal behavior processes. 1983; 9(3): 248–265. [PubMed: 6886630]

Bradlow AR, Bent T. Perceptual adaptation to non-native speech. Cognition. 2008; 106(2):707–729. [PubMed: 17532315]

Brenner N, Bialek W, de Ruyter Van Steveninck R. Adaptive Rescaling Maximizes Information Transmission. Neuron. 2000; 26(3):695–702. [PubMed: 10896164]

Bricker PD, Pruzansky S. Effects of Stimulus Content and Duration on Talker Identification. The Journal of the Acoustical Society of America. 1966; 40(6):1441. [PubMed: 5975580]

Chambers KE, Onishi KH, Fisher C. A vowel is a vowel: generalizing newly learned phonotactic constraints to new contexts. Journal of Experimental Psychology: Learning, Memory, and Cognition. 2010; 36(3):821–828.

Chang F, Dell GS, Bock K. Becoming syntactic. Psychological Review. 2006; 113(2):234–272. [PubMed: 16637761]

Chen S, Gopalakrishnan P. Speaker, environment and channel change detection and clustering via the Bayesian Information Criterion. Proc. darpa broadcast news transcription and understanding workshop. 1998:8.

Cho RY, Nystrom LE, Brown ET, Jones AJ, Braver TS, Holmes PJ, Cohen JD. Mechanisms underlying dependencies of performance on stimulus history in a two-alternative forced-choice task. Cognitive, Affective, & Behavioral Neuroscience. 2002; 2(4):283–299.

Clark A. Whatever next? Predictive brains, situated agents, and the future of cognitive science. Behavioral and Brain Sciences. 2013; 36(3):181–204. [PubMed: 23663408]

Clarke CM, Garrett MF. Rapid adaptation to foreign-accented English. The Journal of the Acoustical Society of America. 2004; 116(6):3647. [PubMed: 15658715]

Clarke-Davidson CM, Luce PA, Sawusch JR. Does perceptual learning in speech reflect changes in phonetic category representation or decision bias? Perception & Psychophysics. 2008; 70(4):604–618. [PubMed: 18556922]

Clayards MA, Tanenhaus MK, Aslin RN, Jacobs Ra. Perception of speech reflects optimal use of probabilistic speech cues. Cognition. 2008; 108(3):804–809. [PubMed: 18582855]

Connine CM, Blasko DG, Hall M. Effects of subsequent sentence context in auditory word recognition: Temporal and linguistic constrainst. Journal of Memory and Language. 1991; 30(2): 234–250.

Courville AC, Daw ND, Touretzky DS. Bayesian theories of conditioning in a changing world. Trends in cognitive sciences. 2006; 10(7):294–300. [PubMed: 16793323]

Creel SC, Aslin RN, Tanenhaus MK. Heeding the voice of experience: the role of talker variation in lexical access. Cognition. 2008; 106(2):633–664. [PubMed: 17507006]

Creel SC, Bregman MR. How Talker Identity Relates to Language Processing. Language and Linguistics Compass. 2011; 5(5):190–204.

Cutler, A.; Eisner, F.; McQueen, JM.; Norris, D. How abstract phonemic categories are necessary for coping with speaker-related variation. In: Fougeron, C.; Kühnert, B.; D'Imperio, M.; Vallée, N., editors. Laboratory phonology. Vol. 10. Berlin: De Gruyter Mouton; 2010. p. 91-111.

Dahan D. The Time Course of Interpretation in Speech Comprehension. Current Directions in Psychological Science. 2010; 19(2):121–126.

Dahan D, Drucker SJ, Scarborough Ra. Talker adaptation in speech perception: adjusting the signal or the representations? Cognition. 2008; 108(3):710–718. [PubMed: 18653175]

Dahan D, Magnuson JS, Tanenhaus MK, Hogan EM. Subcategorical mismatches and the time course of lexical access: Evidence for lexical competition. Language and Cognitive Processes. 2001; 16(5–6):507–534.

Davis MH, Johnsrude IS, Hervais-Adelman A, Taylor K, McGettigan C. Lexical information drives perceptual learning of distorted speech: evidence from the comprehension of noise-vocoded

sentences. Journal of Experimental Psychology: General. 2005; 134(2):222–241. [PubMed: 15869347]

Delattre PC, Liberman AM, Cooper FS. Acoustic Loci and Transitional Cues for Consonants. The Journal of the Acoustical Society of America. 1955; 27(4):769.

Dell GS, Chang F. The P-chain: relating sentence production and its disorders to comprehension and acquisition. Philosophical transactions of the Royal Society of London. Series B, Biological sciences. 2014; 369(1634):20120394. [PubMed: 24324238]

Dell GS, Reed KD, Adams DR, Meyer AS. Speech errors, phonotactic constraints, and implicit learning: A study of the role of experience in language production. Journal of Experimental Psychology: Learning, Memory, and Cognition. 2000; 26(6):1355–1367.

Demberg V, Keller F. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. Cognition. 2008; 109(2):193–210. [PubMed: 18930455]

Dillon B, Dunbar E, Idsardi W. A single-stage approach to learning phonological categories: insights from Inuktitut. Cognitive science. 2013; 37(2):344–377. [PubMed: 23137418]

Drager K. Sociophonetic Variation in Speech Perception. Language and Linguistics Compass. 2010; 4(7):473–480.

Eimas PD, Corbit JD. Selective adaptation of linguistic feature detectors. Cognitive Psychology. 1973; 4(1):99–109.

Eisner F, McQueen JM. The specificity of perceptual learning in speech processing. Perception & Psychophysics. 2005; 67(2):224–238. [PubMed: 15971687]

Eisner F, McQueen JM. Perceptual learning in speech: Stability over time. The Journal of the Acoustical Society of America. 2006; 119(4):1950–1953. [PubMed: 16642808]

Elman J. Finding structure in time. Cognitive Science. 1990; 14:179–211.

Ernestus M. Acoustic reduction and the roles of abstractions and exemplars in speech processing. Lingua. 2014; 142:27–41.

Ernst MO, Banks MS. Humans integrate visual and haptic information in a statistically optimal fashion. Nature. 2002; 415(6870):429–433. [PubMed: 11807554]

Fairhall AL, Lewen GD, Bialek W, de Ruyter Van Steveninck RR. Efficiency and ambiguity in an adaptive neural code. Nature. 2001; 412(6849):787–792. [PubMed: 11518957]

Farmer TA, Brown M, Tanenhaus MK. Prediction, explanation, and the role of generative models in language processing. Behavioral and Brain Sciences. 2013; 36(3):211–212. [PubMed: 23663410]

Feldman NH, Griffiths TL, Goldwater S, Morgan JL. A role for the developing lexicon in phonetic category acquisition. Psychological Review. 2013

Feldman NH, Griffiths TL, Morgan JL. The influence of categories on perception: explaining the perceptual magnet effect as optimal statistical inference. Psychological Review. 2009; 116(4):752–782. [PubMed: 19839683]

Feldman NH, Myers EB, White KS, Griffiths TL, Morgan JL. Word-level information influences phonetic learning in adults and infants. Cognition. 2013; 127(3):427–438. [PubMed: 23562941]

Fine AB, Jaeger T, Farmer TA, Qian T. Rapid Expectation Adaptation during Syntactic Comprehension. PLoS ONE, (Accepted). 2013

Fine AB, Qian T, Jaeger TF, Jacobs RA. Syntactic adaptation in language comprehension. Proceedings of the 1st acl workshop on cognitive modeling and computational linguistics. 2010:18–26.

Frank MC, Goodman ND. Predicting pragmatic reasoning in language games. Science. 2012; 336(6084):998. [PubMed: 22628647]

Frank MC, Goodman ND, Tenenbaum JB. Using speakers' referential intentions to model early cross-situational word learning. Psychological science. 2009; 20(5):578–585. [PubMed: 19389131]

Frank SL, Bod R. Insensitivity of the human sentence-processing system to hierarchical structure. Psychological science. 2011; 22(6):829–834. [PubMed: 21586764]

Friston KJ. A theory of cortical responses. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences. 2005; 360(1456):815–836. [PubMed: 15937014]

Gallistel CR, Mark TA, King AP, Latham PE. The rat approximates an ideal detector of changes in rates of reward: Implications for the law of effect. Journal of Experimental Psychology: Animal Behavior Processes. 2001; 27(4):354. [PubMed: 11676086]

Garnsey S, Pearlmutter N, Myers E, Lotocky M. The Contributions of Verb Bias and Plausibility to the Comprehension of Temporarily Ambiguous Sentences. Journal of Memory and Language. 1997; 37(37):58–93.

Gauvain J-L, Lee C-H. Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains. IEEE Transactions on Speech and Audio Processing. 1994; 2(2): 291–298.

Gelman, A.; Carlin, JB.; Stern, HS.; Rubin, DB. Bayesian Data Analysis. Second. Taylor & Francis; 2003.

Gibson BR, Rogers TT, Zhu X. Human semi-supervised learning. Topics in Cognitive Science. 2013; 5(1):132–172. [PubMed: 23335577]

Goldinger SD. Words and voices: Episodic traces in spoken word identification and recognition memory. Journal of Experimental Psychology: Learning, Memory, and Cognition. 1996; 22(5): 1166–1183.

Goldinger SD. Echoes of echoes? An episodic theory of lexical access. Psychological Review. 1998; 105(2):251–279. [PubMed: 9577239]

Goldinger SD. A complementary-systems approach to abstract and episodic speech perception. Icphs xvi. 2007:49–54.

Goldrick M. Phonological features and phonotactic constraints in speech production. Journal of Memory and Language. 2004; 51(4):586–603.

Gómez R, Gerken L. Infant artificial language learning and language acquisition. Trends in cognitive sciences. 2000; 4(5):178–186. [PubMed: 10782103]

Goodman ND, Stuhlmüller A. Knowledge and implicature: modeling language understanding as social cognition. Topics in cognitive science. 2013; 5(1):173–184. [PubMed: 23335578]

Griffiths TL, Chater N, Kemp C, Perfors A, Tenenbaum JB. Probabilistic models of cognition: exploring representations and inductive biases. Trends in Cognitive Sciences. 2010; 14(8):357–364. [PubMed: 20576465]

Griffiths TL, Sanborn AN, Canini KR, Navarro DJ. Categorization as nonparametric Bayesian density estimation. The probabilistic mind: Prospects for Bayesian cognitive science. 2008:303–328.

Griffiths TL, Vul E, Sanborn aN. Bridging Levels of Analysis for Probabilistic Models of Cognition. Current Directions in Psychological Science. 2012; 21(4):263–268.

Grodner, D.; Sedivy, JC. The effect of speaker-specific information on pragmatic inferences. In: Gibson, EA.; Pearlmutter, NJ., editors. The processing and acquisition of reference. Cambridge, MA: MIT Press; 2011. p. 239-272.

Grosjean F. The recognition of words after their acoustic offset: Evidence and implications. Perception & Psychophysics. 1985; 38(4):299–310. [PubMed: 3831907]

Guediche S, Blumstein SE, Fiez Ja, Holt LL. Speech perception under adverse conditions: insights from behavioral, computational, and neuroscience research. Frontiers in systems neuroscience. 2014 Jan.7:126. [PubMed: 24427119]

Gutfreund Y. Stimulus-specific adaptation, habituation and change detection in the gaze control system. Biological cybernetics. 2012

Hale, J. Second meeting of the north american chapter of the association for computational linguistics on language technologies. Vol. 2. Association for Computational Linguistics; 2001. A probabilistic earley parser as a psycholinguistic model; p. 1-8.

Hanulíková A, van Alphen PM, van Goch MM, Weber A. When one person's mistake is another's standard usage: the effect of foreign accent on syntactic processing. Journal of Cognitive Neuroscience. 2012; 24(4):878–887. [PubMed: 21812565]

Harris H, Gliksberg M, Sagi D. Generalized perceptual learning in the absence of sensory adaptation. Current biology. 2012; 22(19):1813–1817. [PubMed: 22921366]

Hay J, Drager K. Stuffed toys and speech perception. Linguistics. 2010; 48(4):865–892.

Hay J, Warren P, Drager K. Factors influencing speech perception in the context of a merger-in-progress. Journal of Phonetics. 2006; 34(4):458–484.

Heid S, Hawkins S. An acoustical study of long-domain /r/ and /l/ coarticulation. Proceedings of the 5th seminar on speech production: Models and data. 2000:77–80.

Hillenbrand J, Getty LA, Clark MJ, Wheeler K. Acoustic characteristics of American English vowels. The Journal of the Acoustical Society of America. 1995; 97(5.1):3099–3111. [PubMed: 7759650]

Hinton GE. Learning multiple layers of representation. Trends in Cognitive Sciences. 2007; 11(10): 428–434. [PubMed: 17921042]

Huang J, Holt LL. Listening for the norm: adaptive coding in speech categorization. Frontiers in Psychology. 2012 Feb.3:10. [PubMed: 22347198]

Huang Y, Rao RPN. Predictive coding. Wiley Interdisciplinary Reviews: Cognitive Science. 2011; 2(5):580–593. [PubMed: 26302308]

Idemaru K, Holt LL. Word recognition reflects dimension-based statistical learning. Journal of Experimental Psychology: Human Perception and Performance. 2011; 37(6):1939–1956. [PubMed: 22004192]

Jacobs RA. What determines visual cue reliability? Trends in Cognitive Sciences. 2002; 6(8):345–350. [PubMed: 12140085]

Jaeger TF. Categorical Data Analysis: Away from ANOVAs (transformation or not) and towards Logit Mixed Models. Journal of memory and language. 2008; 59(4):434–446. [PubMed: 19884961]

Jaeger TF. Production preferences cannot be understood without reference to communication. Frontiers in Psychology. 2013 Apr.4:230. [PubMed: 23637690]

Jaeger TF, Ferreira VS. Seeking predictions from a predictive framework. Behavioral and Brain Sciences. 2013; 36(4):359–360. [PubMed: 23789872]

Jaeger TF, Snider NE. Alignment as a consequence of expectation adaptation: syntactic priming is affected by the prime's prediction error given both prior and recent experience. Cognition. 2013; 127(1):57–83. [PubMed: 23354056]

Johnson K. The auditory/perceptual basis for speech segmentation. OSU Working Papers in Linguistics. 1997a; 50:101–113.

Johnson, K. Speech perception without speaker normalization: An exemplar model. In: Johnson; Mullennix, editors. Talker variability in speech processing. San Diego: Academic Press; 1997b. p. 145-165.

Johnson K. Resonance in an exemplar-based lexicon: The emergence of social identity and phonology. Journal of Phonetics. 2006; 34(4):485–499.

Johnson K. Factors that affect phonetic adaptation: Exemplar filters and sound change. Talk presented at current issues and methods in speaker adaptation. 2013

Johnson K, Strand Ea, D'Imperio M. Auditory-visual integration of talker gender in vowel perception. Journal of Phonetics. 1999; 27(4):359–384.

Jongman A, Wayland R, Wong S. Acoustic characteristics of English fricatives. The Journal of the Acoustical Society of America. 2000; 108(3):1252. [PubMed: 11008825]

Kamide Y. Learning individual talkers' structural preferences. Cognition. 2012; 124(1):66–71. [PubMed: 22498776]

Kaschak MP, Glenberg AM. This construction needs learned. Journal of Experimental Psychology: General. 2004; 133(3):450–467. [PubMed: 15355149]

Kersten D, Mamassian P, Yuille A. Object perception as Bayesian inference. Annual Review of Psychology. 2004; 55:271–304.

Kleinschmidt, DF.; Fine, AB.; Jaeger, TF. A belief-updating model of adaptation and cue combination in syntactic comprehension. In: Miyake, N.; Peebles, D.; Cooper, RP., editors. Proceedings of the 34th annual conference of the cognitive science society. Austin, TX: Cognitive Science Society; 2012. p. 599-604.Talk

Kleinschmidt DF, Jaeger TF. A re-evaluation of selective adaptation. Manuscript in preparation, University of Rochester, Rochester, NY. 2013

Knill DC, Saunders JA. Do humans optimally integrate stereo and texture information for judgments of surface slant? Vision Research. 2003; 43(24):2539–2558. [PubMed: 13129541]

Knudsen EI. Capacity for Plasticity in the Adult Owl Auditory System Expanded by Juvenile Experience. Science. 1998; 279(5356):1531–1533. [PubMed: 9488651]

Kohn A. Visual adaptation: physiology, mechanisms, and functional benefits. Journal of Neurophysiology. 2007; 97(5):3155–3164. [PubMed: 17344377]

Körding KP, Beierholm U, Ma WJ, Quartz S, Tenenbaum JB, Shams L. Causal inference in multisensory perception. PLoS ONE. 2007; 2(9):e943. [PubMed: 17895984]

Körding KP, Tenenbaum JB, Shadmehr R. The dynamics of memory as a consequence of optimal adaptation to a changing body. Nature Neuroscience. 2007; 10(6):779–786. [PubMed: 17496891]

Kraljic T, Samuel AG. Perceptual learning for speech: Is there a return to normal? Cognitive Psychology. 2005; 51(2):141–178. [PubMed: 16095588]

Kraljic T, Samuel AG. Generalization in perceptual learning for speech. Psychonomic Bulletin & Review. 2006; 13(2):262–268. [PubMed: 16892992]

Kraljic T, Samuel AG. Perceptual adjustments to multiple speakers. Journal of Memory and Language. 2007; 56(1):1–15.

Kraljic T, Samuel AG, Brennan SE. First impressions and last resorts: how listeners adjust to speaker variability. Psychological Science. 2008; 19(4):332–338. [PubMed: 18399885]

Kronrod, Y.; Coppess, E.; Feldman, NH. A Unified Model of Categorical Effects in Consonant and Vowel Perception. In: Miyake, N.; Peebles, D.; Cooper, RP., editors. Proceedings of the 34th annual conference of the cognitive science society. Austin, TX: Cognitive Science Society; 2012. p. 629-634.

Kurumada, C.; Brown, M.; Bibyk, S.; Pontillo, DF.; Tanenhaus, MK. Incremental processing in the pragmatic interpretation of contrastive prosody. In: Knauff, M.; Pauen, M.; Sebanz, N.; Wachsmuth, I., editors. Proceedings of the 35th annual conference of the cognitive science society. Austin, TX: Cognitive Science Society; 2013. p. 846-851.

Kurumada, C.; Brown, M.; Bibyk, S.; Pontillo, F.; Tanenhaus, MK. Rapid adaptation in online pragmatic interpretation of contrastive prosody. In: Bello, P.; Guarini, M.; McShane, M.; Scassellati, B., editors. Proceedings of the 36th annual meeting of the cognitive science society. Austin, TX: Cognitive Science Society; 2014. p. 791-796.

Kurumada, C.; Brown, M.; Tanenhaus, MK. Pragmatic interpretation of contrastive prosody : It looks like speech adaptation. In: Miyake, N.; Peebles, D.; Cooper, RP., editors. Proceedings of the 34th annual conference of the cognitive science society. Austin, TX: Cognitive Science Society; 2012. p. 647-652.

Labov, W. Sociolinguistic Patterns. University of Pennsylvania Press; 1972.

Lancia L, Winter B. The interaction between competition, learning, and habituation dynamics in speech perception. Laboratory Phonology. 2013; 4(1):221–258.

Leggetter C, Woodland P. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. Computer Speech & Language. 1995; 9(2):171–185.

Levy R. Expectation-based syntactic comprehension. Cognition. 2008a; 106(3):1126–1177. [PubMed: 17662975]

Levy, R. A noisy-channel model of rational human sentence comprehension under uncertain input; Proceedings of the 2008 conference on empirical methods in natural language processing; 2008b. p. 234-243.

Levy R, Bicknell K, Slattery T, Rayner K. Eye movement evidence that readers maintain and act on uncertainty about past linguistic input. Proceedings of the National Academy of Sciences of the United States of America. 2009; 106(50):21086–21090. [PubMed: 19965371]

Liberman AM, Cooper FS, Shankweiler DP, Studdert-Kennedy M. Perception of the speech code. Psychological Review. 1967; 74(6):431–461. [PubMed: 4170865]

Linkenhoker BA, von der Ohe CG, Knudsen EI. Anatomical traces of juvenile learning in the auditory system of adult barn owls. Nature neuroscience. 2005; 8(1):93–98. [PubMed: 15608636]

Lisker L, Abramson A. A cross-language sudy of voicing in initial stops: Acoustical measurements. Word. 1964; 20(3):384–422.

Ma WJ, Beck JM, Latham PE, Pouget A. Bayesian inference with probabilistic population codes. Nature Neuroscience. 2006; 9(11):1432–1438. [PubMed: 17057707]

MacDonald MC. How language production shapes language form and comprehension. Frontiers in Psychology. 2013; 4:226. [PubMed: 23637689]

MacDonald MC, Pearlmutter NJ, Seidenberg MS. Lexical nature of syntactic ambiguity resolution. Psychological Review. 1994; 101(4):676–703. [PubMed: 7984711]

MacKay, DJC. Information Theory, Inference, and Learning Algorithms. 3rd. Cambridge, UK: Cambridge University Press; 2003.

Magnuson JS, Nusbaum HC. Acoustic differences, listener expectations, and the perceptual accommodation of talker variability. Journal of Experimental Psychology: Human Perception and Performance. 2007; 33(2):391–409. [PubMed: 17469975]

Marr, D. Vision: A computational investigation into the human representation and processing of visual information. New York, NY, USA: Henry Holt and Co., Inc.; 1982.

Marslen-Wilson W, Warren P. Levels of perceptual representation and process in lexical access: Words, phonemes, and features. Psychological Review. 1994; 101(4):653–675. [PubMed: 7984710]

Massaro, DW. Speech Perception by Ear and Eye: A Paradigm for Psychological Inquiry. Hillsdale, NJ: Erlbaum Associates; 1987.

Massaro, DW. From multisensory integration to talking heads and language learning. In: Calvert, G.; Spence, C.; Stein, BE., editors. The handbook of multisensory processes. Cambridge, MA: MIT Press; 2004. p. 153-176.

Maye J, Aslin RN, Tanenhaus M. The Weckud Wetch of the Wast: Lexical Adaptation to a Novel Accent. Cognitive Science. 2008; 32(3):543–562. [PubMed: 21635345]

Maye J, Werker JF, Gerken L. Infant sensitivity to distributional information can affect phonetic discrimination. Cognition. 2002; 82(3):B101–B111. [PubMed: 11747867]

McClelland JL, Elman JL. The TRACE model of speech perception. Cognitive Psychology. 1986; 18(1):1–86. [PubMed: 3753912]

McGurk H, MacDonald J. Hearing lips and seeing voices. Nature. 1976; 264(5588):746–748. [PubMed: 1012311]

McLennan CT, Luce PA, Charles-Luce J. Representation of lexical form. Journal of Experimental Psychology: Learning, Memory, and Cognition. 2003; 29(4):539–553.

McMurray B, Aslin RN, Toscano JC. Statistical learning of phonetic categories: insights from a computational approach. Developmental Science. 2009; 12(3):369–378. [PubMed: 19371359]

McMurray B, Jongman A. What information is necessary for speech categorization? Harnessing variability in the speech signal by integrating cues computed relative to expectations. Psychological Review. 2011; 118(2):219–246. [PubMed: 21417542]

McMurray B, Tanenhaus MK, Aslin RN. Gradient effects of within-category phonetic variation on lexical access. Cognition. 2002; 86(2):B33–B42. [PubMed: 12435537]

McQueen JM, Cutler A, Norris D. Phonological Abstraction in the Mental Lexicon. Cognitive Science. 2006; 30(6):1113–1126. [PubMed: 21702849]

McQueen JM, Norris D, Cutler A. Lexical influence in phonetic decision making: Evidence from subcategorical mismatches. Journal of Experimental Psychology: Human Perception and Performance. 1999; 25(5):1363–1389.

Miller JL, Connine CM, Schermer TM, Kluender KR. A possible auditory basis for internal structure of phonetic categories. The Journal of the Acoustical Society of America. 1983; 73(6):2124–2133. [PubMed: 6875098]

Mirman D, McClelland JL, Holt LL. An interactive Hebbian account of lexically guided tuning of speech perception. Psychonomic bulletin & review. 2006; 13(6):958–965. [PubMed: 17484419]

Mitterer H, Reinisch E. No delays in application of perceptual learning in speech recognition: Evidence from eye tracking. Journal of Memory and Language. 2013; 69(4):527–545.

Munson B. The Acoustic Correlates of Perceived Masculinity, Perceived Femininity, and Perceived Sexual Orientation. Language and Speech. 2007; 50(1):125–142. [PubMed: 17518106]

Munson B. The influence of actual and imputed talker gender on fricative perception, revisited (L). The Journal of the Acoustical Society of America. 2011; 130(5):2631–2634. [PubMed: 22087888]

Munson, CM. Unpublished doctoral dissertation. University of Iowa; 2011. Perceptual learning in speech reveals pathways of processing.

Neal RM. Slice Sampling. Annals of statistics. 2003; 31(3):705–741.

Nearey TM. Speech perception as pattern recognition. The Journal of the Acoustical Society of America. 1997; 101(6):3241–3254. [PubMed: 9193041]

Nearey TM, Assman PF. Modeling the role of inherent spectral change in vowel identification. The Journal of the Acoustical Society of America. 1986; 80(5):1297.

Newman RS, Clouse Sa, Burnham JL. The perceptual consequences of within-talker variability in fricative production. The Journal of the Acoustical Society of America. 2001; 109(3):1181–1196. [PubMed: 11303932]

Niedzielski N. The Effect of Social Information on the Perception of Sociolinguistic Variables. Journal of Language and Social Psychology. 1999; 18(1):62–85.

Norris D. Shortlist: a connectionist model of continuous speech recognition. Cognition. 1994; 52(3): 189–234.

Norris D, McQueen JM. Shortlist B: A Bayesian model of continuous speech recognition. Psychological Review. 2008; 115(2):357–95. [PubMed: 18426294]

Norris D, McQueen JM, Cutler a. Merging information in speech recognition: feedback is never necessary. The Behavioral and brain sciences. 2000; 23(3):299–325. discussion 325–70. [PubMed: 11301575]

Norris D, McQueen JM, Cutler A. Perceptual learning in speech. Cognitive Psychology. 2003; 47(2): 204–238. [PubMed: 12948518]

Nygaard LC, Pisoni DB. Talker-specific learning in speech perception. Perception & Psychophysics. 1998; 60(3):355–376. [PubMed: 9599989]

Oden GC, Massaro DW. Integration of featural information in speech perception. Psychological Review. 1978; 85(3):172–191. [PubMed: 663005]

O'Donnell, TJ.; Snedeker, J.; Tenenbaum, JB.; Goodman, ND. Productivity and reuse in language. In: Carlson, L.; Hoelscher, C.; Shipley, TF., editors. Proceedings of the 33rd annual conference of the cognitive science society. Austin, TX: Cognitive Science Society; 2011. p. 1613-1618.

Pajak B, Fine AB, Kleinschmidt DF, Jaeger TF. Learning additional languages as hierarchical inference: Insights from L1 processing. Language Learning. 2014 *Submitted*.

Pajak, B.; Levy, R. Phonological Generalization from Distributional Evidence. In: Carlson, L.; Hoelscher, C.; Shipley, TF., editors. Proceedings of the 33rd annual conference of the cognitive science society. Austin, TX: Cognitive Science Society; 2011. p. 2673-2378.

Palmeri TJ, Goldinger SD, Pisoni DB. Episodic encoding of voice attributes and recognition memory for spoken words. Journal of Experimental Psychology: Learning, Memory, and Cognition. 1993; 19(2):309–328.

Pardo, JS.; Remez, RE. The Perception of Speech. In: Traxler, M.; Gernsbacher, MA., editors. The handbook of psycholinguisitcs. 2nd. New York: 2006. p. 201-248.

Perfors A, Tenenbaum JB, Regier T. The learnability of abstract syntactic principles. Cognition. 2011; 118(3):306–338. [PubMed: 21186021]

Peterson GE. Control Methods Used in a Study of the Vowels. The Journal of the Acoustical Society of America. 1952; 24(2):175.

Pickering MJ, Garrod S. An integrated theory of language production and comprehension. Behavioral and Brain Sciences. 2013; 36(4):329–347. [PubMed: 23789620]

Pierrehumbert JB. Word-specific phonetics. Laboratory Phonology. 2002:1–24.

Pierrehumbert JB. Phonetic diversity, statistical learning, and acquisition of phonology. Language and Speech. 2003; 46(Pt 2–3):115–154. [PubMed: 14748442]

Pierrehumbert JB. The next toolkit. Journal of Phonetics. 2006; 34(4):516–530.

Pisoni, DB.; Levi, SV. Representations and representational specificity in speech perception and spoken word recognition. In: Gaskell, MG., editor. The oxford handbook of psycholinguistics. Oxford: Oxford University Press; 2007. p. 3-18.

Pisoni DB, Tash J. Reaction times to comparisons within and across phonetic categories. Perception & Psychophysics. 1974; 15(2):285–290. [PubMed: 23226881]

Qian T, Jaeger TF, Aslin RN. Learning to represent a multi-context environment: more than detecting changes. Frontiers in Psychology. 2012 Jul.3:228. [PubMed: 22833727]

Qian T, Jaeger TF, Aslin RN. Implicit Learning of Bundles of Statistical Patterns in an Incremental Task. Manuscript submitted for publication. 2013

Rao RPN, Ballard DH. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nature Neuroscience. 1999; 2(1):79–87. [PubMed: 10195184]

Reinisch E, Holt LL. Lexically guided phonetic retuning of foreign-accented speech and its generalization. Journal of experimental psychology. Human perception and performance. 2014; 40(2):539–555. [PubMed: 24059846]

Remez RE, Fellowes JM, Rubin PE. Talker identification based on phonetic information. Journal of Experimental Psychology: Human Perception and Performance. 1997; 23(3):651–666. [PubMed: 9180039]

Roberts M, Summerfield Q. Audiovisual presentation demonstrates that selective adaptation in speech perception is purely auditory. Perception & Psychophysics. 1981; 30(4):309–314. [PubMed: 7322807]

Saffran JR, Aslin R, Newport E. Statistical learning by 8-month-old infants. Science. 1996; 274(5294): 1926–1928. [PubMed: 8943209]

Saffran JR, Newport EL, Aslin RN. Word Segmentation: The Role of Distributional Cues. Journal of Memory and Language. 1996; 35(4):606–621.

Saldaña HM, Rosenblum LD. Selective adaptation in speech perception using a compelling audiovisual adaptor. The Journal of the Acoustical Society of America. 1994; 95(6):3658–3661. [PubMed: 8046153]

Salverda AP, Kleinschmidt DF, Tanenhaus MK. Immediate effects of anticipatory coarticulation in spoken-word recognition. Journal of Memory and Language. 2014; 71(1):145–163. [PubMed: 24511179]

Samuel AG. Red herring detectors and speech perception: in defense of selective adaptation. Cognitive Psychology. 1986; 18(4):452–499. [PubMed: 3769426]

Samuel AG. Knowing a Word Affects the Fundamental Perception of The Sounds Within it. Psychological Science. 2001; 12(4):348–351. [PubMed: 11476105]

Samuel AG, Kat D. Early levels of analysis of speech. Journal of Experimental Psychology: Human Perception and Performance. 1996; 22(3):676.

Sanborn AN, Griffiths TL, Navarro DJ. Rational approximations to rational models: Alternative algorithms for category learning. Psychological Review. 2010; 117(4):1144–1167. [PubMed: 21038975]

Sawusch JR. Peripheral and central processes in selective adaptation of place of articulation in stop consonants. The Journal of the Acoustical Society of America. 1977; 62(3):738–750. [PubMed: 903514]

Sharpee TO, Sugihara H, Kurgansky AV, Rebrik SP, Stryker MP, Miller KD. Adaptive filtering enhances information transmission in visual cortex. Nature. 2006; 439(7079):936–942. [PubMed: 16495990]

Sheffert SM, Pisoni DB, Fellowes JM, Remez RE. Learning to recognize talkers from natural, sinewave, and reversed speech samples. Journal of Experimental Psychology: Human Perception and Performance. 2002; 28(6):1447–1469. [PubMed: 12542137]

Shi L, Griffiths TL, Feldman NH, Sanborn AN. Exemplar models as a mechanism for performing Bayesian inference. Psychonomic Bulletin & Review. 2010; 17(4):443–64. [PubMed: 20702863]

Shinoda K, Lee C-H. A structural Bayes approach to speaker adaptation. IEEE Transactions on Speech and Audio Processing. 2001; 9(3):276–287.

Sidaras SK, Alexander JED, Nygaard LC. Perceptual learning of systematic variation in Spanish-accented speech. The Journal of the Acoustical Society of America. 2009; 125(5):3306–3316. [PubMed: 19425672]

Sidaras SK, Nygaard LC. Illusory Vocal Accommodation as a Function of Expected Age of the Speaker. Manuscript submitted for publication. 2014

Simmons JP, Nelson LD, Simonsohn U. False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. Psychological Science. 2011

Sissons HT, Miller RR. Spontaneous recovery of excitation and inhibition. Journal of experimental psychology. Animal behavior processes. 2009; 35(3):419–426. [PubMed: 19594286]

Smith NJ, Levy R. The effect of word predictability on reading time is logarithmic. Cognition. 2013; 128(3):302–319. [PubMed: 23747651]

Smits R. Hierarchical categorization of coarticulated phonemes: A theoretical analysis. Perception & Psychophysics. 2001; 63(7):1109–1139. [PubMed: 11766939]

Sonderegger, M.; Yu, A. A rational account of perceptual compensation for coarticulation. In: Ohlsson, S.; Catrambone, R., editors. Proceedings of the 32nd annual conference of the cognitive science society. Austin, TX: Cognitive Science Society; 2010. p. 375-380.

Staub A, Clifton C. Syntactic prediction in language comprehension: evidence from either…or. Journal of Experimental Psychology: Learning, Memory, and Cognition. 2006; 32(2):425–436.

Staum Casasanto, L. Does Social Information Influence Sentence Processing ?. In: Love, BC.; McRae, K.; Sloutsky, VM., editors. Proceedings of the 30th annual conference of the cognitive science society. Austin, TX: Cognitive Science Society; 2008.

Stocker, AA.; Simoncelli, EP. Sensory Adaptation within a Bayesian Framework for Perception. In: Weiss, Y.; Schölkoph, B.; Platt, J., editors. Advances in neural information processing systems. Vol. 18. Cambridge, MA: MIT Press; 2006. p. 1291-1298.

Strand EA. Uncovering the Role of Gender Stereotypes in Speech Perception. Journal of Language and Social Psychology. 1999; 18(1):86–100.

Strand, EA.; Johnson, K. Gradient and Visual Speaker Normalization in the Perception of Fricatives. In: Gibbons, D., editor. Natural language processing and speech technology: Results of the 3rd konvens conference, bielfelt. Berlin: Mouton de Gruyter; 1996. p. 14-26.

Strange W. Evolving theories of vowel perception. The Journal of the Acoustical Society of America. 1989; 85(5):2081. [PubMed: 2659637]

Sumner M. The role of variation in the perception of accented speech. Cognition. 2011; 119(1):131–136. [PubMed: 21144500]

Tanenhaus MK, Spivey-Knowlton MJ, Eberhard KM, Sedivy JC. Integration of visual and linguistic information in spoken language comprehension. Science. 1995; 268(5217):1632–1634. [PubMed: 7777863]

Tenenbaum JB, Griffiths TL. Generalization, similarity, and Bayesian inference. Behavioral and Brain Sciences. 2001; 24(4):629–640. discussion 652–791. [PubMed: 12048947]

Thanellou A, Green JT. Spontaneous recovery but not reinstatement of the extinguished conditioned eyeblink response in the rat. Behavioral neuroscience. 2011; 125(4):613–625. [PubMed: 21517145]

Toscano JC, McMurray B. Cue integration with categories: Weighting acoustic cues in speech using unsupervised learning and distributional statistics. Cognitive Science. 2010; 34(3):434–464. [PubMed: 21339861]

Toscano JC, Munson CM, Kleinschmidt DF, Jaeger TF. A single mechanism for language acquisition and perceptual learning. under revision. (n.d.).

Trueswell JC, Tanenhaus MK. Semantic influences on parsing: Use of thematic role information in syntactic ambiguity resolution. Journal of Memory and Language. 1994; 33:285–318.

Tunley, A. Unpublished doctoral dissertation. University of Cambridge; 1999. Coarticulatory influences of liquids on vowels in English Alison.

Tzeng CY, Alexander JED, Sidaras SK, Nygaard LC. The Role of Training Structure in Perceptual Learning of Accented Speech. Manuscript submitted for publication. 2014

Vallabha GK, McClelland JL, Pons F, Werker JF, Amano S. Unsupervised learning of vowel categories from infant-directed speech. Proceedings of the National Academy of Sciences of the United States of America. 2007; 104(33):13273–13278. [PubMed: 17664424]

Van Wassenhove V, Grant KW, Poeppel D. Visual speech speeds up the neural processing of auditory speech. Proceedings of the National Academy of Sciences of the United States of America. 2005; 102:1181–1186. [PubMed: 15647358]

van den Bosch, a; Daelemans, W. Implicit Schemata and Categories in Memory-based Language Processing. Language and Speech. 2013; 56(3):309–328. [PubMed: 24416959]

Vatakis A, Spence C. Crossmodal binding: evaluating the "unity assumption" using audiovisual speech stimuli. Perception & psychophysics. 2007; 69(5):744–756. [PubMed: 17929697]

Vitevitch MS, Luce PA. A web-based interface to calculate phonotactic probability for words and nonwords in English. Behavior Research Methods, Instruments, & Computers. 2004; 36(3):481–487.

Vroomen J, van Linden S, de Gelder B, Bertelson P. Visual recalibration and selective adaptation in auditory-visual speech perception: Contrasting build-up courses. Neuropsychologia. 2007; 45(3): 572–577. [PubMed: 16530233]

Vroomen J, van Linden S, Keetels M, de Gelder B, Bertelson P. Selective adaptation and recalibration of auditory speech by lipread information: dissipation. Speech Communication. 2004; 44(1–4): 55–61.

Warker JA, Dell GS. Speech errors reflect newly learned phonotactic constraints. Journal of Experimental Psychology: Learning, Memory, and Cognition. 2006; 32(2):387–398.

Webster, MA.; Werner, JS.; Field, DJ. Adaptation and the Phenomenology of Perception. In: Clifford, C.; Rhodes, G., editors. Fitting the mind to the world: Adaptation and after-effects in high-level vision (advances in visual cognition). Vol. 2. Oxford University Press; 2005. p. 241-277.

Wells JB, Christiansen MH, Race DS, Acheson DJ, MacDonald MC. Experience and sentence processing: statistical learning and relative clause comprehension. Cognitive Psychology. 2009; 58(2):250–271. [PubMed: 18922516]

Whalen DH. Subcategorical phonetic mismatches slow phonetic judgments. Perception & psychophysics. 1984; 35(1):49–64. [PubMed: 6709474]

White KS, Aslin RN. Adaptation to novel accents by toddlers. Developmental Science. 2011; 14(2): 372–384. [PubMed: 21479106]

Wonnacott E, Newport EL, Tanenhaus MK. Acquiring and processing verb argument structure: distributional learning in a miniature language. Cognitive psychology. 2008; 56(3):165–209. [PubMed: 17662707]

Yildirim, I.; Degen, J.; Tanenhaus, MK.; Jaeger, TF. Linguistic Variability and Adaptation in Quantifier Meanings. In: Knauff, M.; Pauen, M.; Sebanz, N.; Wachsmuth, I., editors. Proceedings of the 35th annual conference of the cognitive science society. Austin, TX: Cognitive Science Society; 2013. p. 3835-3840.

Zhu, X.; Rogers, T.; Qian, R.; Kalish, C. Humans Perform Semi-Supervised Classification Too; Proceedings of the 22nd aaai conference on artificial intelligence; 2007. p. 864-869.

Appendix

Modeling methods and assumptions

This appendix lays out the basic Bayesian belief updating model proposed as part of the ideal adapter framework. We first summarize the formal specification of the model. As we describe in the main text, the model assumes that adaptation takes place over a cue dimension that might integrate auditory and visual information. The arguments for this assumption, potential caveats, and our reply to those caveats are summarized next. We then describe how the model's parameters were fit to the behavioral data from Vroomen et al. (2007) and our own study. Throughout these sections we state the assumptions made by our model and model fitting procedure. Crucially, none of these assumptions creates a bias in favor of our hypothesis. Finally, we provide a table that summarizes our assumptions.

Model specification

To quantify and test the qualitative predictions of the ideal adapter framework, we implemented a basic Bayesian belief updating model. This model makes a number of simplifying assumptions. The most notable is that we consider only two categories in this

model, /b/ and /d/, and assume that the prior beliefs about their means and variances are independent

$$p(\boldsymbol{\mu},\boldsymbol{\sigma}^2)=\prod_c p(\mu_c,\sigma_c^2)=p(\mu_\mathrm{b},\sigma_\mathrm{b}^2,\mu_\mathrm{d},\sigma_\mathrm{d}^2) \quad (26)$$

Combined with the fact that participants in these experiments reliably classify even the auditorily ambiguous audio-visual adaptors as the intended categories (Vroomen et al., 2004), this simplifies belief updating (Equation 8).

The form of the individual category parameter priors was also chosen for reasons of computational convenience. We use the conjugate prior for the Normal distribution with unknown mean and variance, a Normal-χ^{-2} distribution (Gelman et al., 2003). This distribution factorizes the joint prior into two components:

$$p(\mu_c,\sigma_c^2)=p(\mu_c|\sigma_c^2)p(\sigma_c^2) \quad (27)$$

$$=\mathrm{Normal}(\mu_c|\mu_{0,c},\sigma_c^2/\kappa_0)\chi^{-2}(\sigma_c^2|\nu_0,\sigma_{c,0}^2) \quad (28)$$

The prior on the variance is a χ^{-2} distribution, with two parameters, ν_0 and $\sigma_{c,0}^2$. $\sigma_{c,0}^2$ is the expected value of the category variance σ_c^2, while ν_0 represents the effective prior sample size for the mean (reflecting the uncertainty about $\sigma_{c,0}^2$, see main text) The prior on the mean is conditioned on the value of the category variance, and is a Normal distribution. The expected value of that normal distribution is the prior mean parameter $\mu_{0,c}$. Its variance is σ_c^2/κ_0, i.e., the *category* variance divided by the effective prior sample size for the mean κ_0.

Belief updating with a conjugate prior—Using a conjugate prior is convenient because after updating with observations $X = x_1, \ldots, x_n$ (whose mean value is $\bar{x}$ and sample variance is $s^2=\frac{1}{n}\sum_i(x_i-\overline{x})^2$), the posterior is *also* a Normal-χ^{-2} distribution, with updated parameters (Gelman et al., 2003):

$$\kappa_n=\kappa_0+n \quad (29)$$

$$\mu_{n,c}=\frac{\kappa_0}{\kappa_n}\mu_{0,c}+\frac{n}{\kappa_n}\overline{x} \quad (30)$$

$$\nu_n=\nu_0+n \quad (31)$$

$$\sigma_{n,c}^2=\frac{1}{\nu_n}(\nu_0\sigma_{c,0}^2+ns^2+\frac{n\kappa_0}{\kappa_n}(\mu_{0,c}-\overline{x})^2) \quad (32)$$

These parameter updates have intuitive interpretations: the κ_0 and ν_0 parameters are 'pseudocounts', or the effective sample size of the prior. To update them from their prior to posterior values, they are both incremented by n, the number of observations. The updated

mean $\mu_{n,c}$ is a weighted average of the sample mean (weighted by the actual sample size n) and the prior mean (weighted by the prior effective sample size of the mean κ_0). The updated expected variance $\sigma^2_{n,c}$ is also a weighted average, of the observed variance ns^2 (weighted by the actual sample size n), the prior expected variance $\sigma^2_{0,c}$ (weighted by the effective prior sample size, ν_0), and a third term, which accounts for deviation of the observed mean from the expected mean $\mu_{c,0}$. This last term is weighted by $n\kappa_0/\kappa_n$, which gets larger (relative to the weights of the other terms, n and ν_0) when $n \approx \kappa_0$.

What happens to the expected mean and variance when more and more observations are made? Specifically, what happens when n becomes much larger than the prior effective sample sizes for the variance ν_0 and mean κ_0? First, the posterior mean μ_n converges to the observed mean x (since $n/\kappa_n = n/(\kappa_0 + n) \approx 1$ when $n \gg \kappa_0$). Second, and similarly, the expected variance converges on the observed variance. That is, with a lot of data about the current situation the model's beliefs will converge against the actual statistics of that situation. In both cases, the stronger the prior beliefs (larger prior sample sizes κ_0 and ν_0), the more observations it takes to overcome the listener's prior beliefs. These prior effective sample size parameters can thus be understood as controlling the strength and speed of adaptation effects.

Incremental belief updating—If all the individual, observed cue values for each category are assumed to be statistically independent (conditioned on the mean and variance), then the joint likelihood of all observed cues is equal to the product of the individual likelihoods, and thus

$$p(\mu_c, \sigma_c^2 | x_1 \dots x_N) \propto \underbrace{p(x_1, \dots, x_N | \mu_c, \sigma_c^2)}_{\text{joint likelihood}} \underbrace{p(\mu_c, \sigma_c^2)}_{\text{prior}} \quad (33)$$

$$\propto \left(\prod_{i=1}^{N} p(x_i | \mu_c, \sigma_c^2) \right) p(\mu_c, \sigma_c^2) \quad (34)$$

$$\propto p(x_N | \mu_c, \sigma_c^2) \left(\prod_{i=1}^{N-1} p(x_i | \mu_c, \sigma_c^2) \right) p(\mu_c, \sigma_c^2) \quad (35)$$

$$\propto \underbrace{p(x_N | \mu_c, \sigma_c^2)}_{\text{likelihood of } x_N} \underbrace{p(\mu_c, \sigma_c^2 | x_1, \dots, x_{N-1})}_{\text{updated prior}} \quad (36)$$

That is, the listener's beliefs after the Nth observation are a combination of their beliefs about the means and variances after all $N - 1$ preceding observations, combined with the likelihood of the current observation given those beliefs.

One insight this provides is that it is not always necessary to maintain a full record of all observations made so far, or even their statistics. Rather, it is sufficient to just track the posterior distribution over means and variances after each token. This leads naturally to

approximate inference methods like particle filters, which approximate beliefs after $N-1$ observations via a set of "particles", each a particular set of means and variances, with the particles collectively approximating the full distribution $p(\mu_c, \sigma_c^2)$. After observation N, each particle's estimate is updated, with particles failing to predict observation N effectively being thrown out and particles which do effectively predict observation N persisting or even being "cloned" to replace the rejected particles. Particle filters have been shown to be a reasonable approximation of Bayes-optimal inference in distributional learning categorization problems, and match human performance well even with a small number of particles (Sanborn et al., 2010).

Audio-visual cue integration

Next, we discuss the model's assumptions about the nature of the dimension over which adaptation takes place. This part of the model is not a consequences of the ideal adapter framework. Instead, it is motivated by evidence for cross-modal interactions in audio-visual speech processing (Bejjanki, Clayards, et al., 2011; McGurk & MacDonald, 1976). Given evidence for such cross-modal interactions, we remain agnostic about the level at which listeners are adapting to the audio-visual stimuli. Specifically, we treat as a free parameter whether listeners adapt to a purely auditory representation of the perceived cues, or to some representation which integrates information from both auditory and visual cues. After describing how this was implemented in the model, we summarize possible objections to this aspect of the model and our reply to these objections.

Linear cue combination—Under reasonably general assumptions, information from auditory and visual cues to the same phonetic dimension can be optimally combined into a multimodal cue value by a weighted sum $x = w_a x^{(a)} + w_\upsilon x^{(\upsilon)}$, where the weights w_a and w_υ sum to 1 and are proportional to the reliability of the auditory and visual cues (Bejjanki, Clayards, et al., 2011; Ernst & Banks, 2002; Jacobs, 2002; Knill & Saunders, 2003; Toscano & McMurray, 2010).

We incorporate this into our model via treating the perceived cue values x as a weighted sum of the continuum values for the auditory and visual tokens $x = wx^{(\upsilon)} + (1-w)x^{(a)}$, where the weight w is a free parameter. For the audio-visual adaptors used in this experiment, in the /b/ condition the visual cue indicates a prototypical /b/, and so $x^{(\upsilon)} = 1$, while the auditory cue indicates an ambiguous /b/-/d/, $x^{(a)} = x_{\text{bd}} \approx 5$ (depending on the particular participant's most ambiguous stimulus).

The linear combination of these two cues results in an integrated cue estimate somewhere in between, not quite prototypical but not fully ambiguous. Note that for our model fits, the best-fitting cue weights are in general roughly equal ($w \approx 0.5$), which suggests that the perceived cue value for the audiovisual adaptor is substantially less ambiguous than the auditory cue. This is hardly surprising given well-known cross-modal effects on speech perception with similar consequences, such as the McGurk effect (McGurk & MacDonald, 1976). The assumption that the audio-visual stimulus is not really ambiguous is also consistent with our finding from pilot studies that participants can reliably classify the ambiguous audiovisual adaptor stimuli (which also justifies somewhat our assumption—

made solely for the sake of convenience—that the category label of each adaptor stimulus is known with certainty).

Is there evidence against audio-visual integration in adaptation?—Our decision to include audio-visual cue integration in the model is supported by a wealth of evidence that audio and visual cues are processed together in speech perception (e.g. Bejjanki, Clayards, et al., 2011; Massaro, 2004; Van Wassenhove, Grant, & Poeppel, 2005; Vatakis & Spence, 2007). There are, however, two studies that have looked specifically at selective adaptation to audio-visual adaptors, and contrary to what we propose here, have concluded that selective adaptation is driven entirely by the adaptor's audio component (Roberts & Summerfield, 1981; Saldaña & Rosenblum, 1994). Since these studies might be taken to argue that we introduce unnecessary complexity into the model, we briefly discuss them.

The audio-visual adaptors used by Roberts and Summerfield (1981) and Saldaña and Rosenblum (1994) differ from the adaptors used by Vroomen et al. (2007) and our replication, in that the audio and visual components had large, categorical mismatches. Roberts and Summerfield (1981) used auditory-/b/, visual-/g/ adaptors, intended to evoke a /d/ percept (as in McGurk & MacDonald, 1976), but found selective adaptation effects on a /b/-/d/ continuum that were indistinguishable from an audio-visual /b/ adaptor. However, their participants did not generally perceive the adaptor in this way, with only half reporting an alveolar /d/ or /ð/, and the others reporting /kl/, /m/, of /fl/. This contrasts with with the perception of the stimuli in Vroomen et al. (2007) and our own studies, where participants reliably classified the audio-visual adaptor stimuli as 'labeled' by the visual component.

Saldaña and Rosenblum (1994) is more relevant for the current purpose. They used an audio-/ba/, visual-/va/ adaptor stimulus which was consistently identified as /va/ by participants. This produced a selective adaptation effect on a /ba/-/va/ continuum equivalent to that of its audio /ba/ component presented separately. However, it is not possible to tell whether the observed effect was due to selective adaptation of /b/, or recalibration of /v/, since both would produce a shift in the category boundary towards /b/. If the visual and auditory cues are integrated as we have tentatively proposed, and the visual cue weight is higher than the auditory weight, then the audio-visual integration would be an imperfect /v/, leading in our model to both a /v/ percept and recalibration of /v/ and fewer /b/ responses.

In sum, it is broadly accepted that speech perception involves cross-modal cue integration. Whether adaptation can take place over those integrated cues is an open question that previous literature does not speak to. Our own results suggest a positive answer (since the best-fitting visual cue weight $w \approx .5$ in all our studies).

The (ir)relevance of audio-visual integration for the interpretation of our results—Treating the adaptor cue value as a linear combination of the audio and visual cues has two main effects in our model. First, it causes recalibration to saturate somewhere below the maximum possible aftereffect of +1 (all adaptor-category responses). Second, it causes recalibration to peak and then decrease with further exposure. Neither of these consequences comes from the combination of audio and visual cues in the model *per se*, but

rather from the fact that such combination causes the adaptor percept to be perceived as not fully ambiguous.

There are other possible reasons this might occur besides cue integration, such as a perceptual magnet effect. Feldman et al. (2009) explain the perceptual magnet effect as a result of the listener's attempt to infer what cue value the talker *intended* to produce based on an *observed* cue that is corrupted by noise variability or sensory uncertainty.[33] In such a model, the listener's knowledge of the category distributions acts as an additional cue, which is combined with the noisy percept. Many of the aspects of the perceptual magnet effect can be explained by assuming that what the listener perceives—for the purposes of making responses in a perceptual magnet experiment—is the best guess about the talker's intended production, which is a combination of the actual cue value perceived and the cue values expected based on the category structure. Furthermore, when the listener knows that the talker intended to produce a particular *category*, this framework predicts that the perceived cue is pulled towards that category mean, which is very similar to the effect of cue integration with a prototypical visual cue value.

Regardless of whether the perceptual magnet effect (through visual labeling of the auditory stimulus) or early audio-visual integration cause the audio-visual stimulus to be substantially less ambiguous than the auditory, the underlying (statistical) logic of the two approaches is very similar. The standard cue integration model (Ernst & Banks, 2002) assumes that the perceiver is trying to get a good estimate of an (unobserved) quantity (like the talker's intended production) which is noisily approximated by multiple cues. When the two cues are corrupted by independent Gaussian noise, this gives rise to an optimal strategy of taking an average of the individual estimates yielded by the two cues, weighted according to their reliability. Similarly, in the perceptual magnet model of Feldman et al. (2009), when the category and noise distributions are Gaussian, the best estimate of the intended cue value is the same, reliability- (inverse variance-) weighted average of the category mean and the observed cue value.[34]

Why would the ideal adapter combine information about the intended production from multiple cues before adaptation? Within-category variability is not only meaningless noise, but rather might additionally reflect factors like coarticulation, or other systematic changes in cue value. This means that the cue values we have treated as relevant only for making a single, categorical decision (e.g. /b/ vs. /d/) are actually potentially informative about other nearby segments as well. Thus, the talker's intended *cue value*, in addition to their intended category, reflects nearby categories as well. This fact, combined with non-uniform

[33]Here we use "noise" as a shorthand for noise and sensory uncertainty. Such uncertainty is not necessarily due to noise in the sense of random variability but also arises from, for instance, the limited resolution in the neural representation of particular stimulus parameters or a mismatch in the type of features encountered and those assumed by upstream neural decoding (Beck, Ma, Pitkow, Latham, & Pouget, 2012).

[34]For the purposes of adaptation, the major practical difference between an explanation in terms of the perceptual magnet effect and an explanation in terms of cross-modal cue integration would be that as the listener updates their beliefs about the category mean, the value of the perceptual magnet cue would change with exposure, whereas the value of the visual cue (presumably) does not. Of relevance is evidence that prolonged repeated exposure can also erase *lexically*-driven recalibration (where there is no visual cue available): Vroomen et al. (2007) re-analyzed data from Samuel (2001), and found that a lexically disambiguated, auditorily ambiguous adaptor elicited the same pattern of initial positive and long-run negative aftereffects as their visually-disambiguated adaptor. This suggests that an explanation of long-run recalibration behavior which relies on adaptation to integration of audio-visual cues per se is inadequate.

transitional probabilities between different categories, means that when considered as a distribution over multiple cues, the distribution for each category can be thought of as highly structured, to the extent that meaningful variability in a cluster of cues is stronger than meaningless noise variability in those cues.

With multiple cues, within-category variability due to the influence of neighboring segments might thus lie a low-dimensional manifold, the shape and structure of which is determined in part by how possible contexts influence the cue values corresponding to each category. Treating the bottom-up sensory signal as a likelihood over intended productions (as do Feldman et al., 2009) and combining it with a structured prior in essence *filters* the uncertain sensory estimates in such a way as to maximize information about meaningful within-category variation in cue values and minimize variability due to uninformative noise.

Using such an integrated cue value for belief updating would not be strictly optimal in a Bayesian sense. It could still, however, reflect constraints imposed by a system which is optimized for processing running speech, rather than estimating the distributions of raw acoustic cue values for each single category in isolation. Support for this idea comes from findings that listeners *do* use information about the onset of an upcoming noun contained in the vowel of the determiner ‘the’ to launch saccades to the corresponding target in a visual world task before the onset of the target noun itself (Salverda, Kleinschmidt, & Tanenhaus, 2014).

Next, we describe how the model was fit to the behavioral data from Vroomen et al. (2007) and our own study.

Model fitting and parameter estimation

The updating rules of the conjugate prior (29) suggest natural ways of fitting the model to the data from Vroomen et al. (2007). First, there is a natural separation between the expected value parameters, $\mu_{c,0}$ and $\sigma^2_{c,0}$, which determine the category means and variances the listener believes are *most* probable before the adaptation phase begins, and the effective prior sample size parameters, κ_0 and ν_0, which determine how willing they are to update those beliefs. The expected means and variances can be set a priori, based on pre-test data. These are thus fixed by the data, rather than being free parameters. Only the effective prior sample size parameters must be fit to the actual adaptation data.

While it is in principle possible to fit each participant’s data individually, the amount of data available from each participant is very small (only six trials per test block) and leads to unstable parameter estimates. We thus chose to use the aggregate data. Another possibility would be to fit a model with a linked prior on the hyperparameters ν_0, κ_0, and *w* that allows for systematically limited variability between listeners. While the insights into possible individual differences would be enlightening, the primary purpose of the current study is to demonstrate the mechanics of the proposed framework. We therefore leave additional modeling improvements to future work.

Estimating prior expected means and variances from pre-test data—We used the pre-test classification data collected by Vroomen et al. (2007) to estimate the underlying means and variances of the corresponding Gaussian mixture (Feldman et al., 2009).

For a mixture of two Gaussian distributions—/b/ and /d/—with equal variance σ^2, the categorization function $p(C = \text{b} \mid x)$ is a logistic function $(1 + \exp(-gx + b))^{-1}$, with slope g and intercept b related to the means μ_b and μ_d and the variance σ^2:

$$g = \frac{\mu_\text{b} - \mu_\text{d}}{\sigma^2} \text{ and } b = \frac{\mu_\text{b}^2 - \mu_\text{d}^2}{\sigma^2} = \frac{(\mu_\text{b} + \mu_\text{d})(\mu_\text{b} - \mu_\text{d})}{\sigma^2} \quad (37)$$

To estimate b and g from the pre-test data, one additional degree of freedom in Equation 37 needs to be held constant. We chose to fix the distance between the means, $\mu_\text{b} - \mu_\text{d}$. Given these values, the values for $(\mu_\text{b} + \mu_\text{d})/2$ (the middle of the participant's subjective continuum) and σ^2 can be calculated using

$$\frac{\mu_\text{b} + \mu_\text{d}}{2} = \frac{b}{g} \text{ and } \sigma^2 = \frac{\mu_\text{b} - \mu_\text{d}}{g} \quad (38)$$

The difference between the means sets the scale of the continuum, and we chose to use $\mu_\text{b} - \mu_\text{d} = 8$, the length (in steps) of the acoustic continuum, which stretches from $x = 1$ (derived from a natural /aba/) to $x = 9$ (from a natural /ada/). This is roughly equivalent to assuming that all subjects would accept these tokens as good productions of /aba/ and /ada/, which indeed they do (Vroomen et al., 2004).

This method makes two assumptions. First, it assumes that participant's subjective prior probabilities of /b/ vs. /d/ (regardless of the cue value) are equal. This is not difficult to relax (it only shifts the boundary of the classification function by the log ratio of the prior probabilities; Feldman et al., 2009), and the model's predictions are qualitatively unchanged when the prior probability of /b/ vs. /d/ is included as a free parameter.

Second, it assumes that the prior variance of the two categories is equal. There are two reasons why this assumption (though probably false) is sufficient for our purpose. First, based on pilot simulations, asymmetric prior variance results in asymmetries between recalibration by x_bd^b and x_bd^d which appear as a overall bias towards more /b/ or /d/ responses with further exposure, regardless of the exposure category. Using the aftereffect difference score as the dependent measure largely removes any effect of this bias, because taking the difference between the /b/ and /d/ exposure conditions removes this positive correlation (see Figure 5). Second, and more importantly, asymmetric prior variance does not change the qualitative predictions about the build-up and decay of recalibration overall, and the purpose of setting prior parameters based on non-adaptation data is to reduce the flexibility of the model in order to more clearly evaluate the hypothesis that phonetic adaptation reflects incremental belief updating.[35]

Individual listeners' classification functions show a fair amount of variability, but all have comparable slopes and boundaries roughly in the middle of the continuum. For this reason,

we fit a mixed effects logistic regression model (Jaeger, 2008) to the pre-test data, which allows for some variability in the slope and intercept of each subject, and so better estimates the slope and intercept most representative of the population. The prior parameters were set based on this slope and intercept as above, with $\mu_{b,0} = 1.10$, $\mu_{d,0} = 9.10$ and $\sigma^2_{b,0}=\sigma^2_{d,0}=3.74$.

Generating predictions from prior confidence and visual cue weight hyperparameters—The free parameters—ν_0, κ_0, and w—were fit to the listeners' responses during test trials, which occurred after exposure to 1,2,4, …, 256 adaptor stimuli.

Model predictions were generated for each test block in the following way. For a test block after n cumulative exposures to the /b/ adaptor, model predictions are generated assuming that the observed values are n repetitions of the most ambiguous cue value ($X = \{x_1, \ldots, x_n\}$, $x_i = x_{bd}$) which are labeled as /b/ with very high certainty ($C = \{c_1, \ldots, c_n\}$, c_i = b) by the visual component (and vice-versa for /d/ exposure trials). The response of the ideal adapter to test stimulus x_{test} depends on the posterior distribution over category parameters given the exposure to the adaptor thus far, $p(\mu_b, \sigma^2_b|X, C)$:

$$p(c_{\text{test}}=\text{b}|x_{\text{test}}, X, C)= \int\int p(c_{\text{test}}=\text{b}, \mu_\text{b}, \sigma^2_\text{b}|x_{\text{test}}, X, C)\text{d}\mu_\text{b}\text{d}\sigma^2_\text{b} \quad (39)$$

$$= \int\int p(c_{\text{test}}=\text{b}|x_{\text{test}}, \mu_\text{b}, \sigma^2_\text{b})p(\mu_\text{b}, \sigma^2_\text{b}|X, C)\text{d}\mu_\text{b}\text{d}\sigma^2_\text{b} \quad (40)$$

Because of the conjugate prior we used, the posterior $p(\mu_b, \sigma^2_b|X, C)$ is found analytically, by updating the hyperparameters (κ_0, ν_0, $\mu_{b,0}$, $\sigma^2_{b,0}$) as described above in equations (29)–(32). Specifically, they were updated with the sample statistics, which have count n, mean $\bar{x} = wx_{visual} + (1 - w)x_{bd}$ (where the visual cue value $x_{visual} = 1$ for visual /b/ and $x_{visual} = 9$ for visual /d/), and sample variance $s^2 = 0$.

The particular choice of prior is also convenient in that the integral in (39) can be evaluated analytically, with the result that the marginal likelihood $p(x_{test} \mid c_{test} = \text{b}, X, C)$ has a scaled t distribution, with mean $\mu_{b,n}$, variance (squared scale) $\frac{(1+\kappa_n)\sigma^2_{b,n}}{\kappa_n}$, and degrees of freedom ν_n (Gelman et al., 2003). The marginal likelihood of the x_{test} under the other, un-adapted category, is analogously found from the prior hyperparameters, κ_0, ν_0, $\mu_{d,0}$, $\sigma^2_{d,0}$, and the marginal posterior probability $p(c_{test} = \text{b} \mid x_{test}, X, C)$ can then be found for each test stimulus based on Equation 40. The marginal posterior probabilities for the three types of test stimuli are averaged to produce the model-predicted probability of a /b/ or /d/ response on test block n (depending on whether the visual cue indicated /b/ or /d/, respectively).

[35]Note that it would in theory be possible to estimate prior category variances directly from other data, removing the need for the simplifying assumption of equal prior variances. For example, prior variances could be estimated from aggregate or even individual production data, discriminability data (Kronrod et al., 2012), goodness-of-exemplar judgments (Pisoni & Tash, 1974; Andruski, Blumstein, & Burton, 1994) or any combination thereof. The exploration of these directions for model improvement are left for future work.

Likelihood of test stimuli—To fit to the overall data, the likelihood of the data was calculated based on the number of responses that were the same category as the adaptor. These adaptor-category response counts were summed in each block. That is, for a /b/-exposure block, /b/ responses were considered 'positive' responses, while /d/ responses were considered 'positive' responses for test blocks during /d/ exposure. This adaptor-category response rate is essentially equivalent to the aftereffect difference score: if y_b and y_d denote the proportion of /b/ responses after /b/ and /d/ exposure respectively, $z_b = y_b$ and $z_d = 1 - y_d$ denote the proportion of adaptor category responses after /b/ and /d/ exposure, and $z = \frac{z_b + z_d}{2}$ is the average adaptor category response overall, then the aftereffect can be found via $2z - 1 = z_b + z_d - 1 = z_b - (1 - z_d) = y_b - y_d = y_{\mathrm{AE}}$.

The likelihood of the adaptor-category response counts given the total number of trials and the model predicted adaptor-category response probability for each block and condition (derived from particular values of the hyperparameters ν_0, κ_0, and w) was evaluated by a binomial likelihood distribution. Specifically, if z_j is the number of adaptor-category responses at test block j, out of n_j test trials in that block (in both cases summing across participants), and θ_j is the model-predicted probability of adaptor-category response, then the likelihood of block j is

$$p(y_j|n_j,\theta_j) = \begin{pmatrix} n_j \\ y_j \end{pmatrix} \theta_j^{y_j}(1-\theta_j)^{n_j - y_j} \quad (41)$$

and the joint likelihood of the data is

$$p(y|n,\theta) = \prod_j p(y_j|n_j,\theta_j) \quad (42)$$

Because a binomial likelihood was used for fitting the model to the data, the error bars on the data show the confidence intervals for the rate parameter of a binomial distribution with the observed counts of adaptor-category responses and non-adaptor-category responses. These were calculated as the 2.5% and 97.5% quantiles of the posterior distribution for the adaptor-category response rate, which is $\mathrm{Beta}(z_j + \frac{1}{2}, n_j - z_j + \frac{1}{2})$ assuming a non-informative (Jeffrey's) $\mathrm{Beta}(\frac{1}{2}, \frac{1}{2})$ prior (Gelman et al., 2003). These quantiles were transformed to the aftereffect scale for visualization in the same way as the data, and for the proportion-/b/ response plots were calculated based on the number of /b/ and /d/ responses instead of the number of adaptor-category responses.

Sampling and hyperpriors—This joint data likelihood was combined with a weak, regularizing prior, $p(\log \kappa_0) = p(\log \nu_0) = \mathrm{Normal}(0, 100)$, which has a 95% interval that stretches from about $\nu_0 = 10^{-9}$ to 10^9, with a mode of 1. Any prior sample size that is a few times larger than the maximum sample size (256) results in essentially no adaptation; in this range the prior is essentially constant (the prior probability of $\nu_0 = 1$, the value which maximizes the prior probability, is only 1.2 times greater than the prior probability of $\nu_0 =$

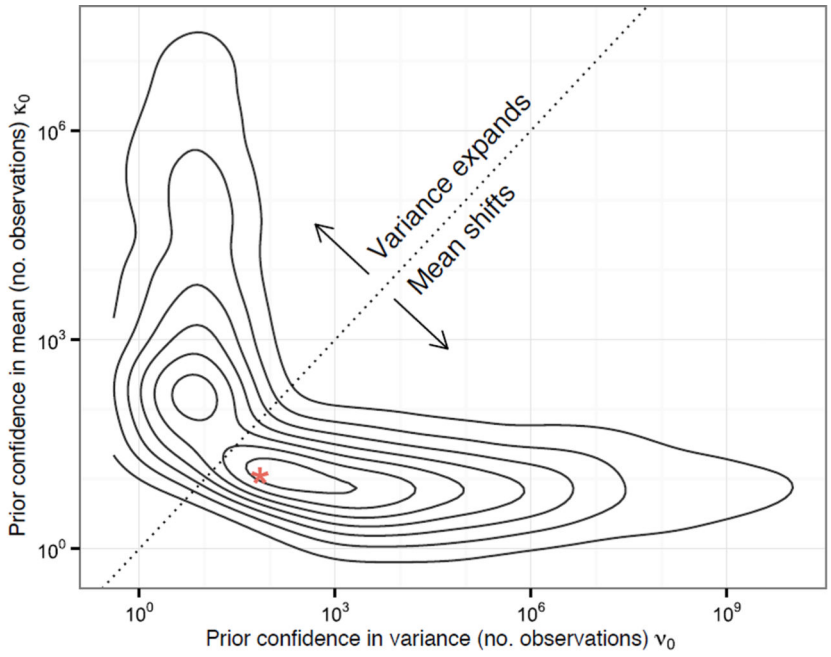
1000, and likewise for κ_0), and thus this prior has essentially no influence on the fit of the model. For the visual cue weight *w* the prior was uniform between 0 and 1 and thus uninformative.

The posterior distribution of the hyperparameters is not easy to find analytically, and so samples were drawn from this distribution using a hybrid Gibbs/slice sampler, where each hyperparameter is sampled in turn via slice sampling (Neal, 2003), given the last sampled values of the other parameters. The samples can be used to find the maximum *a posteriori* (MAP) estimate of the best-fitting parameter values, as well as the full joint posterior. The joint posteriors of the confidence parameters (mean prior pseudocount κ_0 and variance prior pseudocount ν_0) for the fits to the build-up of recalibration, build-up of selective adaptation, long-term effects in both, and the Mechanical Turk replication data are shown in Figure A1.

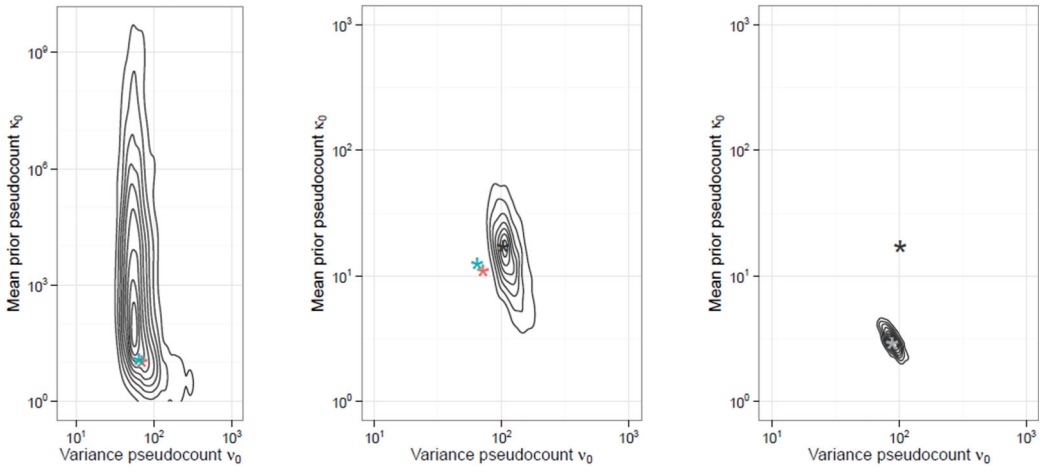
Model assumptions

In order to relate the qualitative predictions of the ideal adapter framework to the behavior of human listeners, it is necessary or convenient to make some simplifying assumptions. Table A1 review these assumptions, whether they are justified, and if not, whether violating them leads to problems for the conclusions we reach from our modeling results. Many of them have already been introduced and discussed above. None of these assumptions bias the results of the modeling towards better fits. If anything many of them make it harder for the model to fit data which violates them.

All of these assumptions are assumptions of the model, and *not* the ideal adapter framework. In particular, the assumption that all observed cues (from one category) are identically distributed goes against the basic point of the ideal adapter analysis that cue distributions change from one situation to another, but it is a necessary simplification for specifically modeling beliefs about cues in the *particular* situation of a laboratory experiment. Also, a true ideal adapter would use their prior experience with category variance and base rate category probabilities to set these for each category, but we assume they are equal for /b/ and /d/ because this is a convenient simplification which reduces the number of hyperparameters that need to be estimate to fit the model without qualitatively changing the predictions.



(a) Vroomen et al. (2007), recal. 64 exposures



(b) V07, sel. ad., 64 (c) V07, simult. (d) MTurk data, simult.

Figure A1. MCMC-estimated joint posterior density (contours) and MAP-estimates (asterisks) of prior confidence parameters for all studies presented in the main text. Black asterisk shows MAP estimate for combined fit, and colored asterisks show earlier fits. Blue asterisk shows the MAP estimate for the selective adaptation condition, and red the recalibration condition. Panel (a): perceptual recalibration data from Vroomen et al. (2007) up 64 exposures (replotted for convenience). Panel (b): selective adaptation data from Vroomen et al. (2007) up to 64 exposures. Panel (c): data from both recalibration and selective adaptation Vroomen et al. (2007) up to 256 exposures. Panel (d): data from our

web-based experiment. Light grey asterisk shows MAP estimate for these fits and dark asterisk shows MAP estimates from fit to Vroomen et al. (2007) data for comparison.

Table A1

Assumptions of the belief updating model used to evaluate the ideal adapter framework predictions.

Assumption	Simplification justified by data?	Problem?
Independently and identically distributed cues.	No: non-stationarity/lack of invariance means cue distributions are different.	Not for modeling first block adaptation in the lab (listeners seem to assume that they're in a totally new situation).
Equal prior variances	Probably not, a priori. Also, pilot simulations show that asymmetrical prior variance leads to an overall increase or decrease in the proportion of /b/ responses *regardless* of the exposure category, and the same pattern shows up in the data.	Not when using the aftereffect measure, which effectively controls for changes in overall rate of /b/ responses.
Equal prior probability of /b/ vs. /d/	No. /b/ is more than twice as likely as /d/ in this context (Vitevitch & Luce, 2004). However, listeners make roughly equal numbers of /b/ and /d/ responses during pre-test so they could infer that /b/ and /d/ are equally likely in this task.	No, based on pilot simulations there's no qualitative difference.
No change in beliefs about prior probabilities	Unclear.	No. When confidence in prior probability is included as a free hyperparameter, it's always inferred to be very high (no change).
Labeled data (supervised adaptation)	Yes. Listeners can classify the ambiguous audio-visual stimuli nearly perfectly (98%), and they can't discriminate between ambiguous and prototypical audio-visual adaptor from the same category (52% on an ABX task; Vroomen et al., 2004).	
Only labeled data counts for adaptation	Unclear. Listeners can adapt to shifted distributions without additional information (C. M. Munson, 2011). Fully optimal ideal adapter predicts beliefs should be partially updated (7), but depends on tracking the full posterior distribution for all previous observations which may be psychologically implausible.	Probably not: doubling the number of test trials doesn't change adaptation. Other category learning studies suggest that when there are many unlabeled training items, they have little influence on later behavior (Zhu et al., 2007), but it's a question for future work.
Adaptation to integrated audio and visual cues	Maybe, see main text for discussion.	Probably not. There are other reasons why the ambiguous audiovisual adaptor might not be perceived as ambiguous for the purposes of adaptation as discussed above.
Normal distributions for cues; Normal-χ^{-2} parameter distributions	Normal cue distributions are a common assumption in computational modeling, and using a conjugate prior is a natural, convenient choice.	No: any distribution that is informationally efficient (has few enough effective parameters) would predict the same kind of rapid/stable adaptation.
Order of trials doesn't matter (exchangeability)	No: Kraljic et al. (2008) and carry over effects between blocks observed in our own data and Vroomen et al. (2007) (see Supplementary Material for discussion).	No: when modeling just the first block of exposure to simple cue statistics, exchangeability is probably reasonable.
Independence of prior beliefs about different categories	No. For instance: vowel F1 means are all higher for female vs. male talkers, which introduces positive correlations between the means across talkers (Hillenbrand et al., 1995).	Probably not, especially when prior beliefs are weak, as we have argued is expected in laboratory studies with unusual speech (and is supported by the estimates of weak prior confidence).

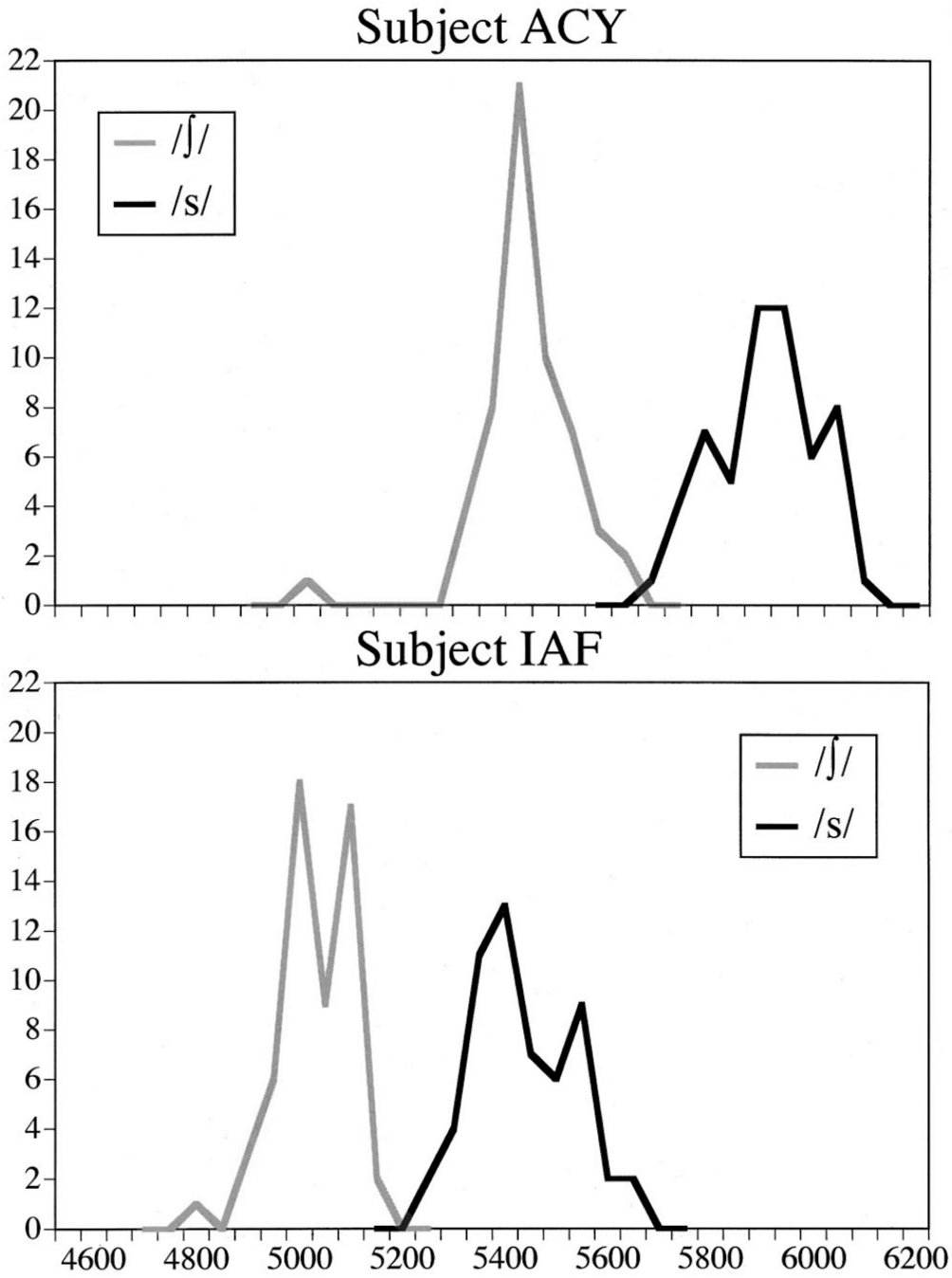


Figure 1.

Distribution of frication frequency centroids, a crucial cue to the contrast between /s/ and /ʃ/, from two talkers (reproduced with permission from Newman et al., 2001, copyright 2001 Acoustical Society of America).

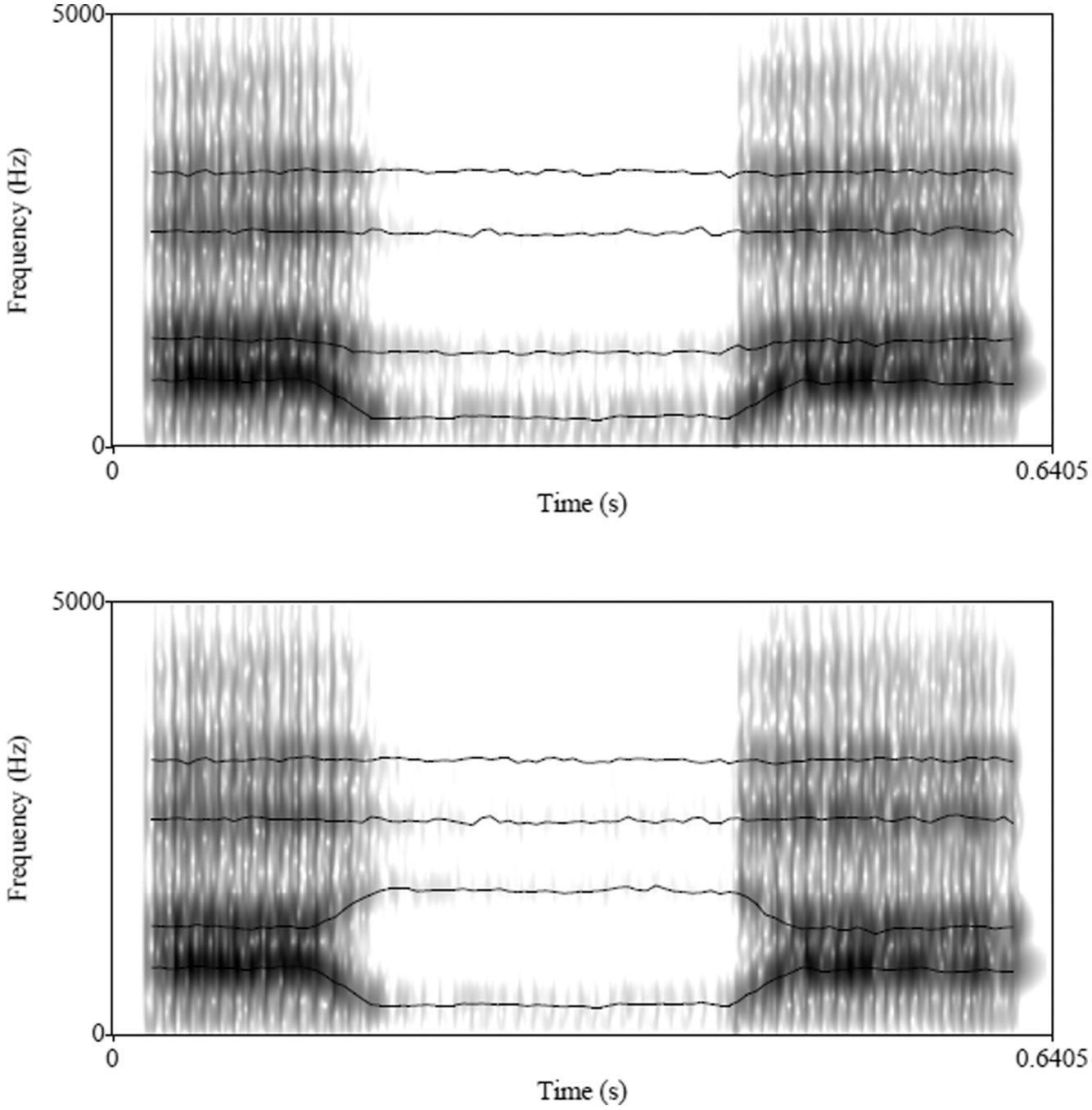


Figure 2.

Spectrograms for /aba/ (top) and /ada/ (bottom) with formant tracks (synthesized as described in Vroomen et al. (2004) and provided by Jean Vroomen). Note the higher second formant (F2) locus for the transitions into and out of the closure for /ada/.

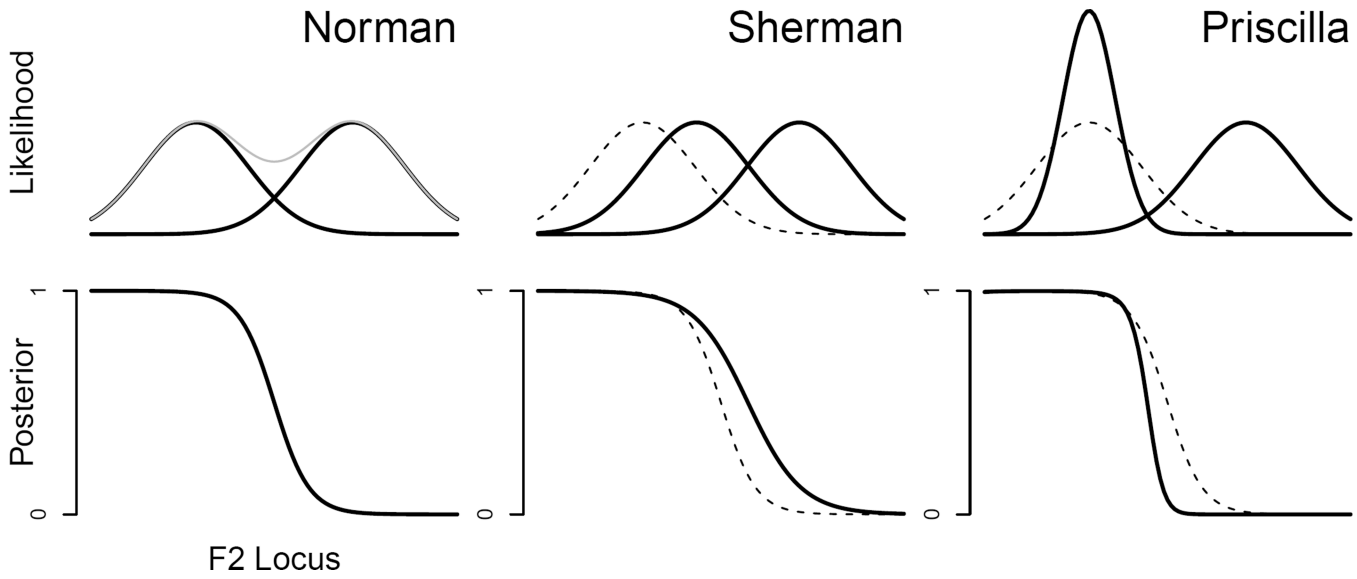


Figure 3.

Relationship between F2 locus likelihood functions $p(x|c)$ (top) and posterior probability of /b/, or classification function $p(c|x)$ (bottom; assuming $p(c) = 0.5$ for both c =/b/ and /d/), for three different talkers: a 'normal' talker (Norman), a 'shifted' talker (Sherman), and a 'precise' talker (Priscilla). Dashed lines show the /b/ likelihood function and classification function corresponding to the 'normal' talker. Light gray line in top left shows the marginal likelihood, $p(x) = \Sigma\, p(x|c)p(c)$, which corresponds to the overall distribution of cue values, regardless of which category they came from, and is the sum of the two likelihood functions.

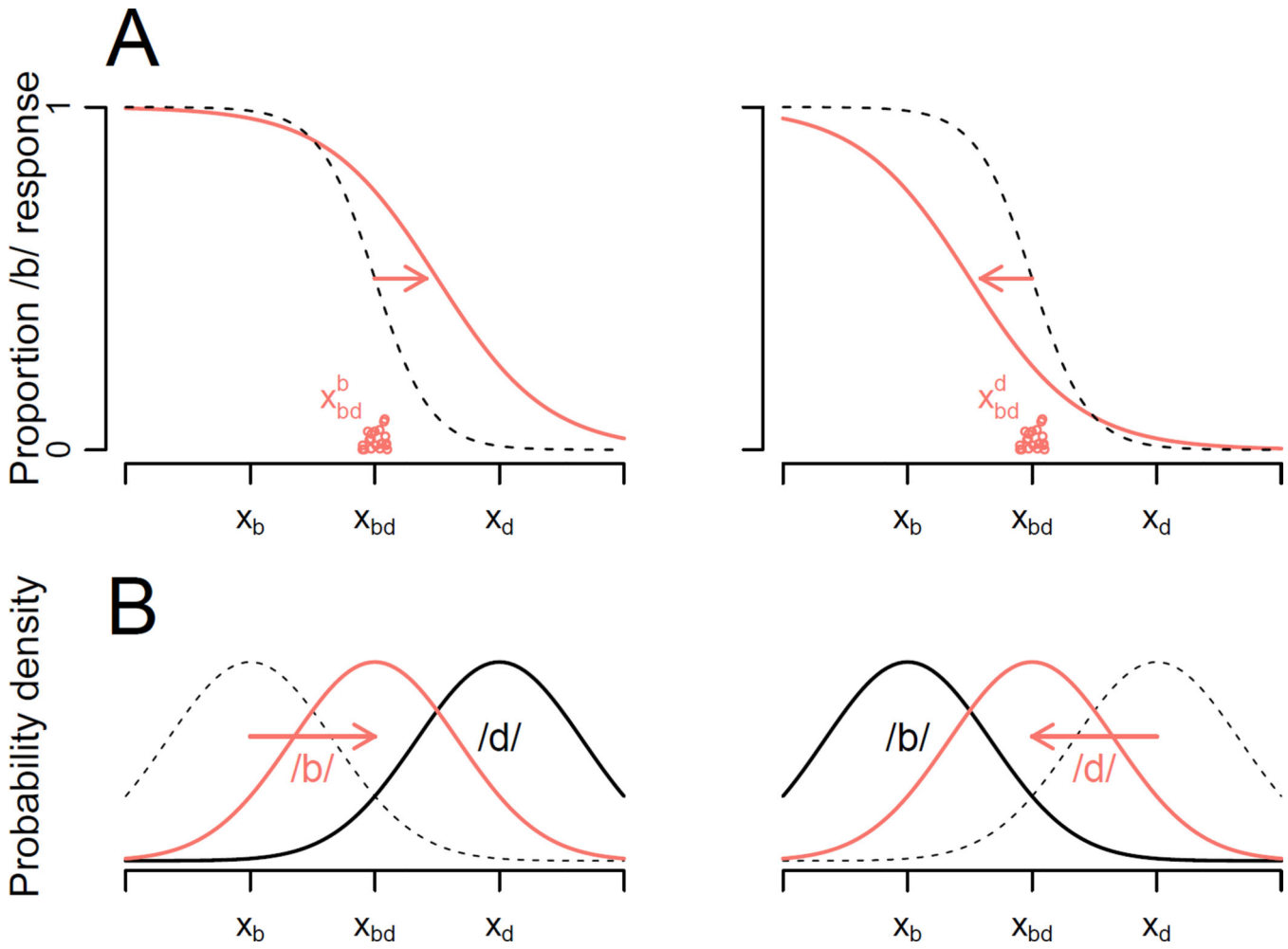


Figure 4.

Schematic illustration of the results of perceptual recalibration on classification of a /b/-/d/ continuum (A), and the changes in the listener's beliefs about the underlying distributions which we propose to account for the changes in classification (B). Dashed lines show pre-exposure classification functions and distributions, while solid lines show post-recalibration. Left panels show the results of exposure to $x^{\mathrm{b}}_{\mathrm{bd}}$, and the right to $x^{\mathrm{d}}_{\mathrm{bd}}$.

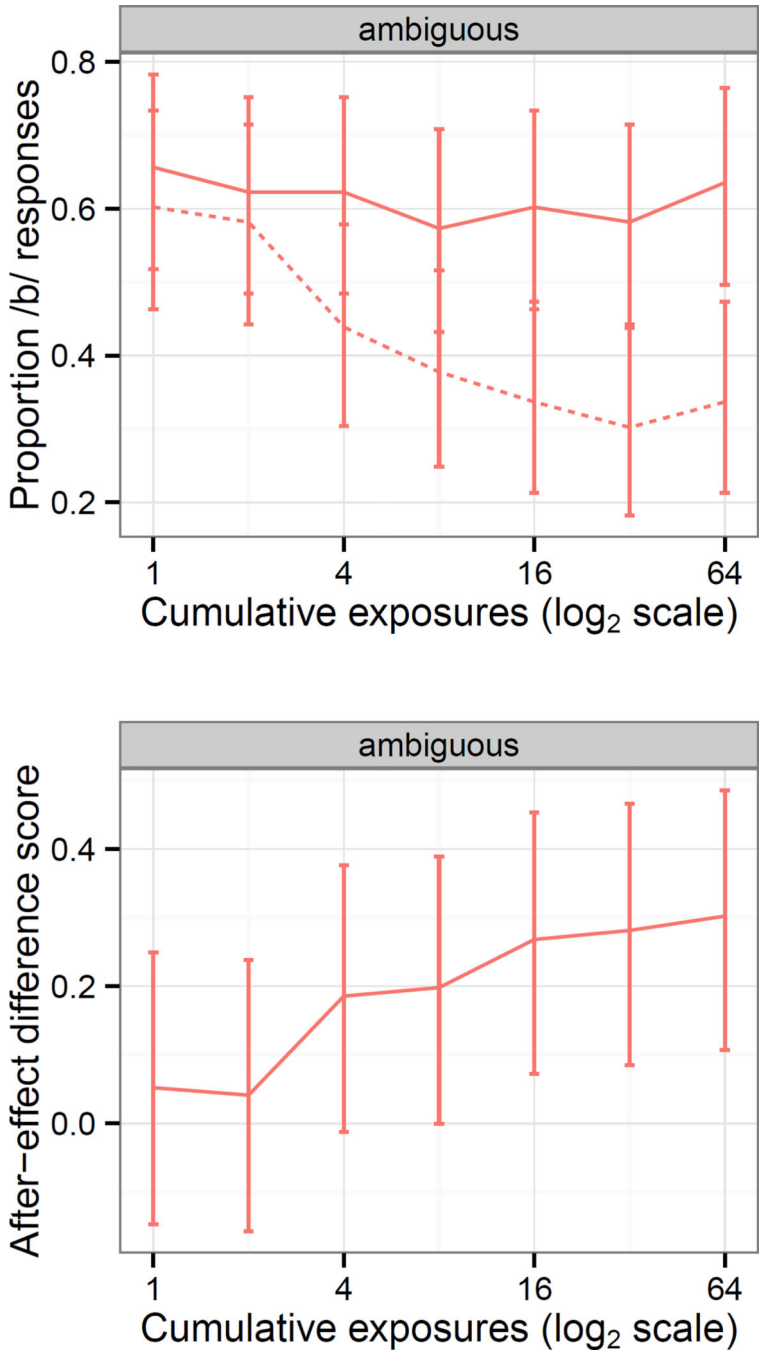


Figure 5.

Recalibration results from Vroomen et al. (2007), showing both the proportion of /b/ responses (top, solid line /b/ exposure and dashed line /d/ exposure) and the aftereffect difference score (bottom) for the first 64 critical exposures in the first exposure block. Error bars indicate 95% confidence intervals.

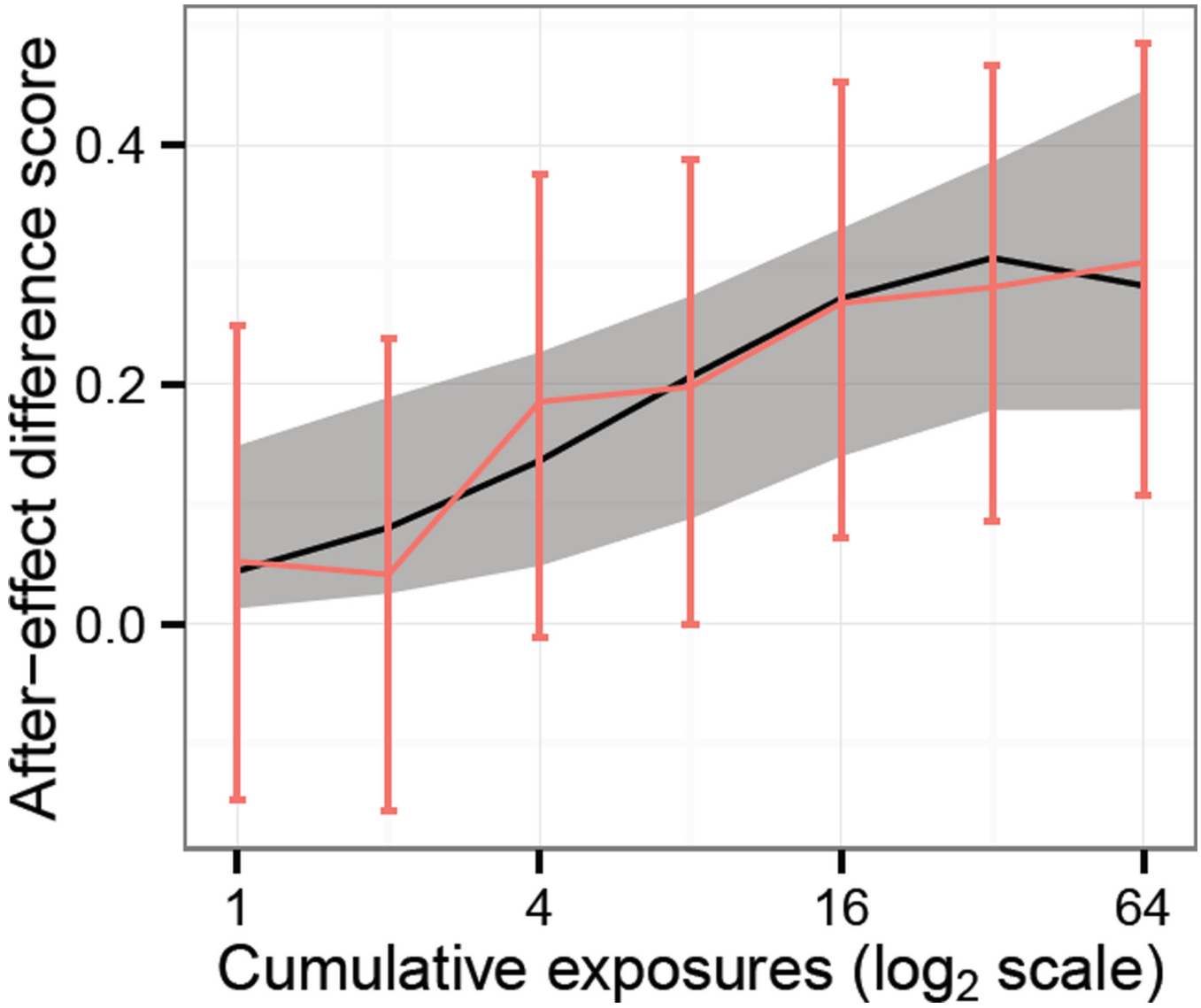


Figure 6.

Belief updating model fit to build-up of recalibration data from Vroomen et al. (2007). The *x*-axis shows the number of cumulative exposures to the adaptor (on a log scale), and the *y*-axis shows the aftereffect difference score. The solid black line shows the MAP (maximum a posteriori) estimate predictions ($r^2 = 0.96$). The error bars and shaded region show 95% credible intervals for the data and model predictions, respectively (see Appendix A).

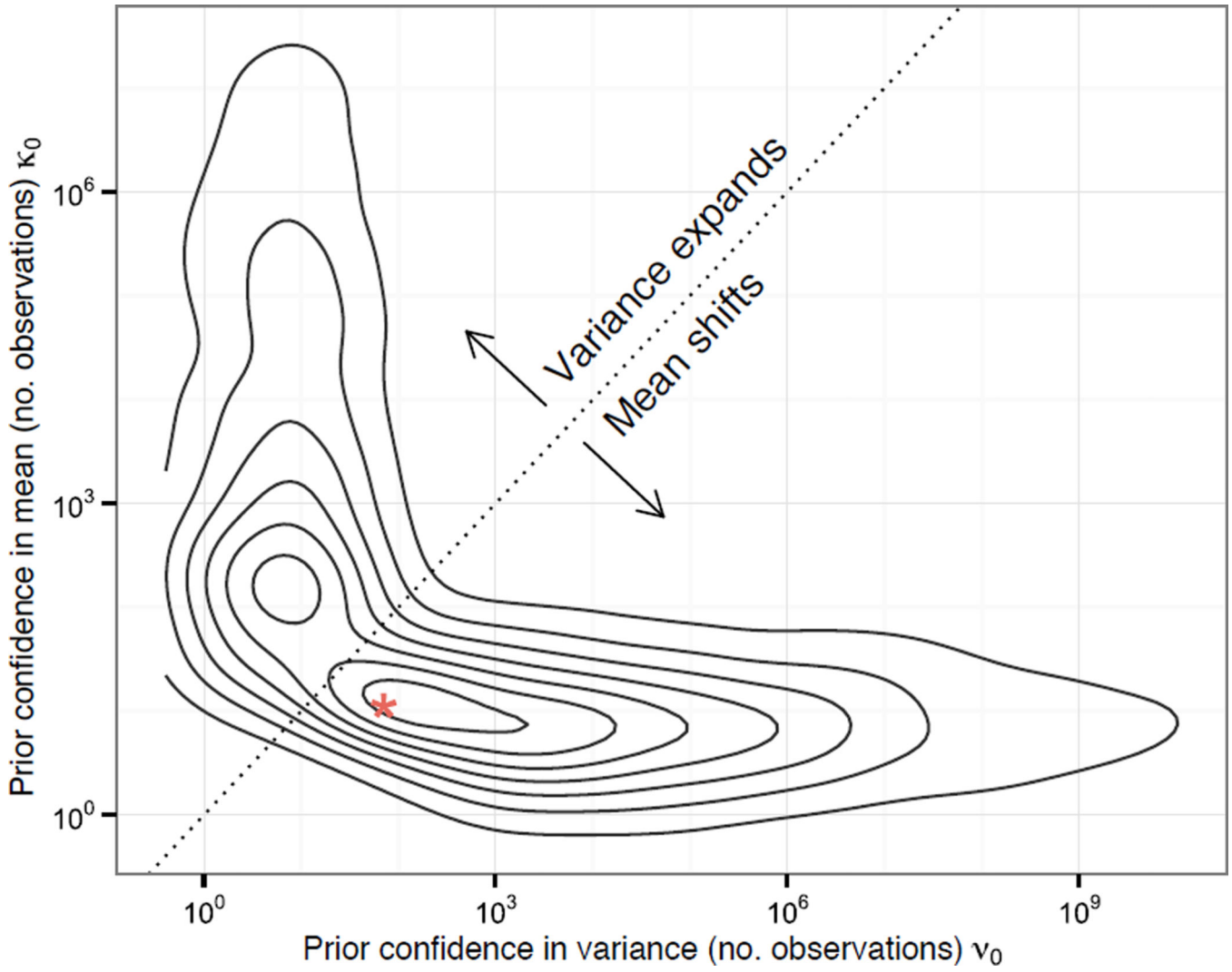


Figure 7.

The belief-updating model finds two different ways of fitting the build-up of recalibration (Figure 6), as illustrated by this density plot of the distribution of the mean and variance prior confidence parameters (κ_0 and ν_0, respectively) that are consistent with the data (estimated by samples via MCMC). The diagonal shows solutions with equal confidence in prior beliefs about the mean and variance. Points below the line have higher confidence in the variance, and adapt by shifting the category mean. Points above the line have higher confidence in the *mean* and adapt by expanding the category variance. Note that even though the best-fitting parameters (red asterisk, and curve in Figure 6) are below the line (and shift the mean), there are areas of high posterior probability on both sides of the line (hills on the contour plot).

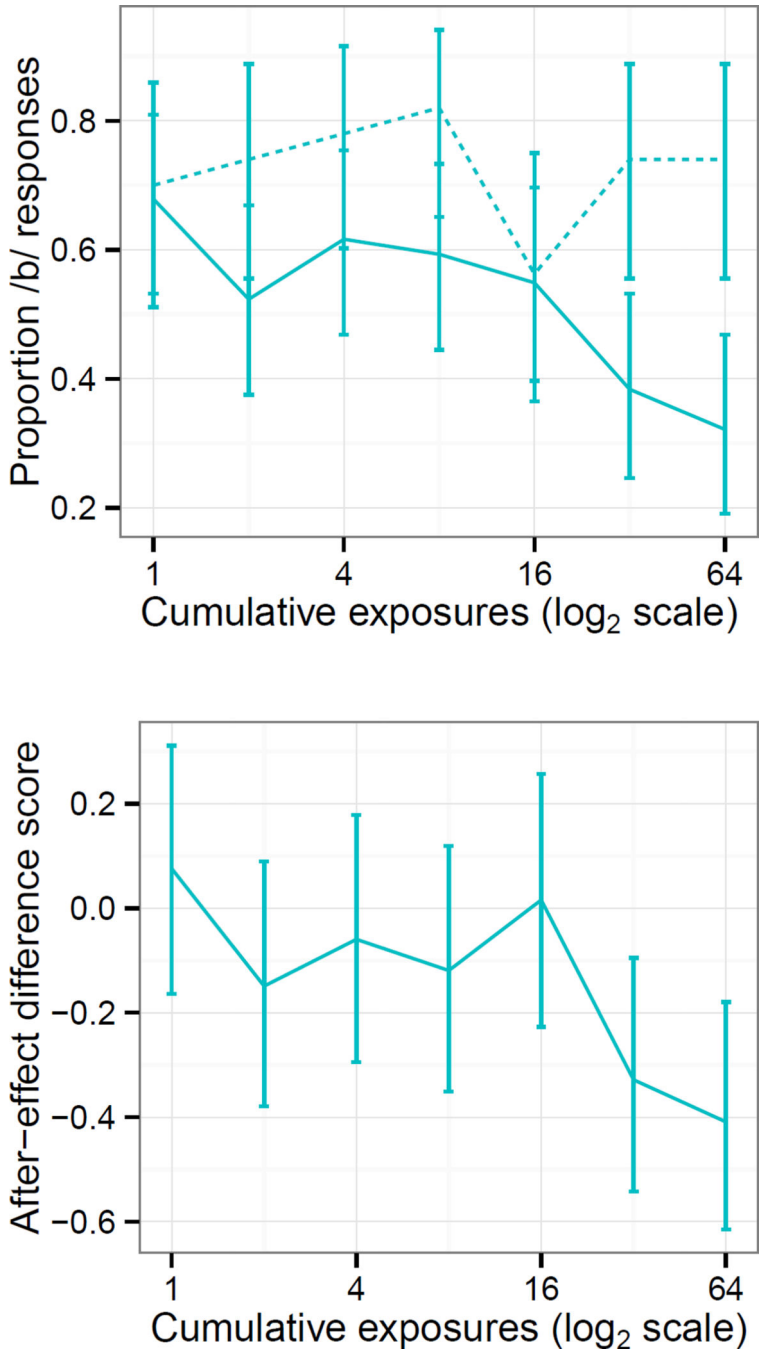


Figure 8.

Selective adaptation results from Vroomen et al. (2007), showing both the proportion of /b/ responses (top) and the aftereffect difference score (bottom) for the first 64 cumulative exposures in the first exposure block. Error bars indicate 95% confidence intervals.

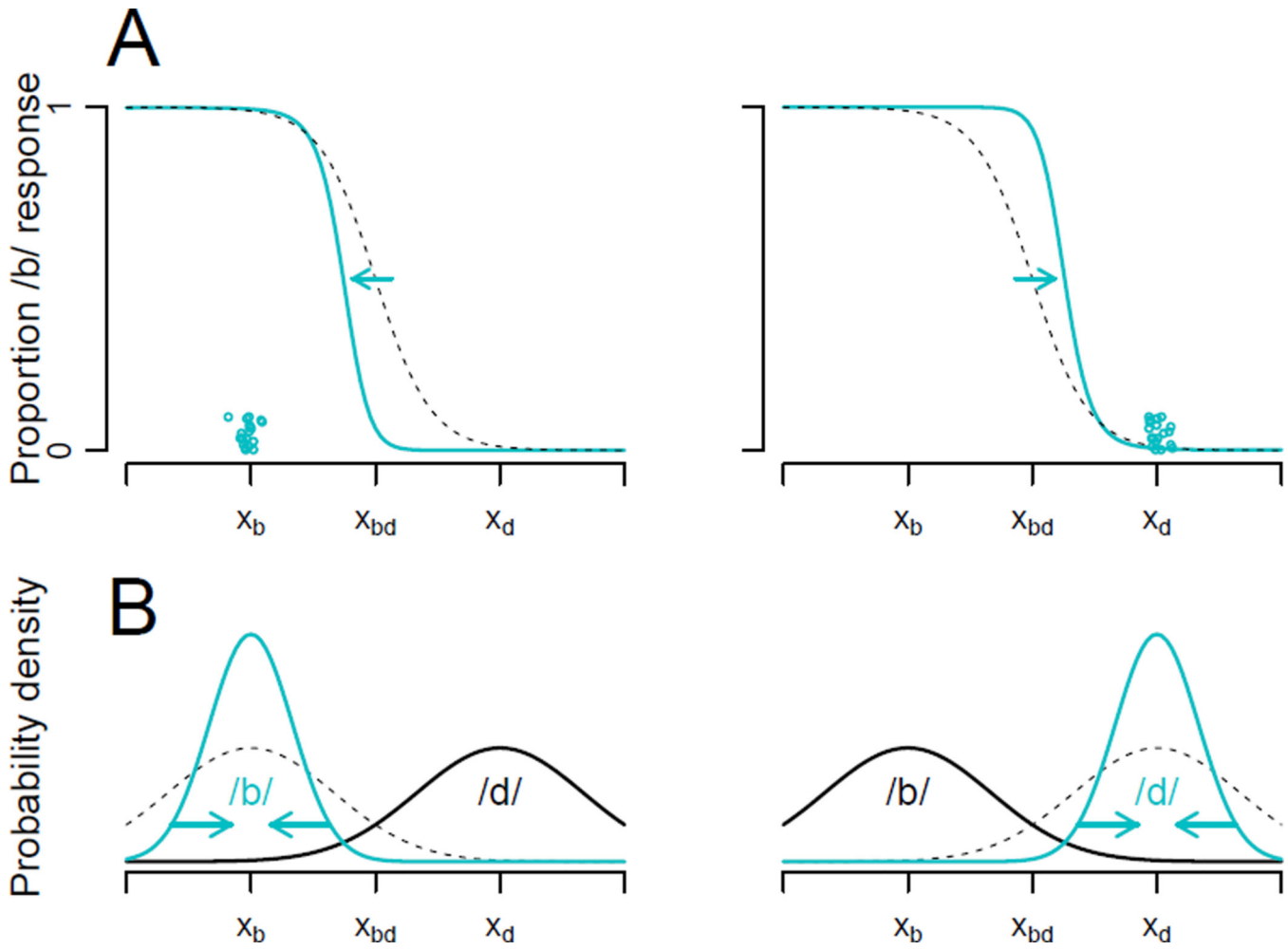


Figure 9.

Schematic illustration of the results of selective adaptation on classification of a /b/-/d/ continuum (A), and the changes in the listener's beliefs about the underlying distributions which we propose to account for the changes in classification (B). Dashed lines show pre-exposure classification functions and distributions, while solid lines show post-recalibration. Left panels show the results of exposure to prototypical $x_{\mathrm{b}}^{\mathrm{b}}$, and the right to $x_{\mathrm{d}}^{\mathrm{d}}$.

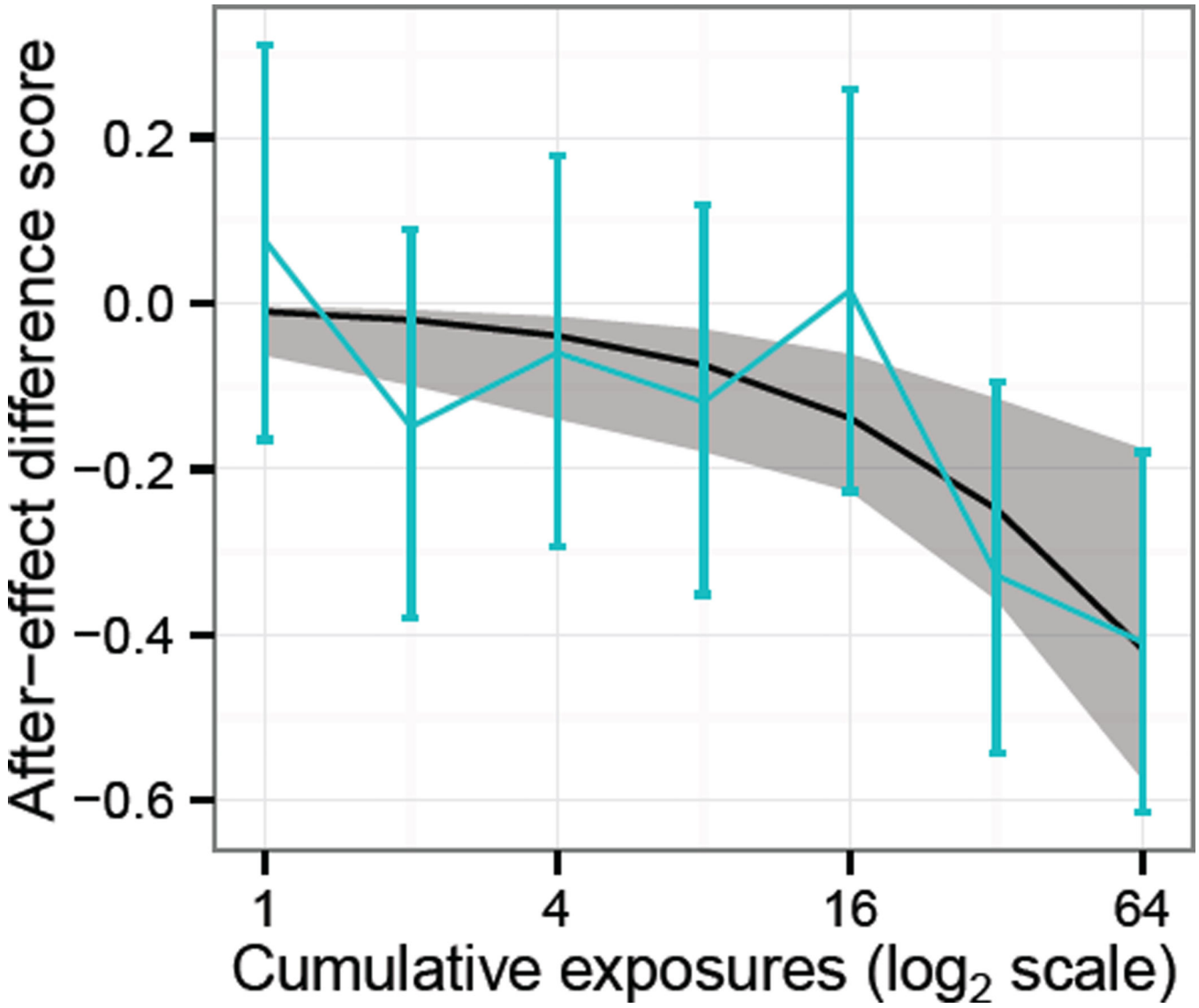


Figure 10.

Belief-updating model fits to the data from Vroomen et al. (2007) on the build up of selective adaptation after varying levels of cumulative exposure to a prototypical audio-visual adaptor (*x*-axis, on a log scale). The solid black line shows the MAP (maximum a posteriori) estimate predictions ($r^2 = 0.83$). The error bars and shaded region show 95% credible intervals for the data and model predictions, respectively.

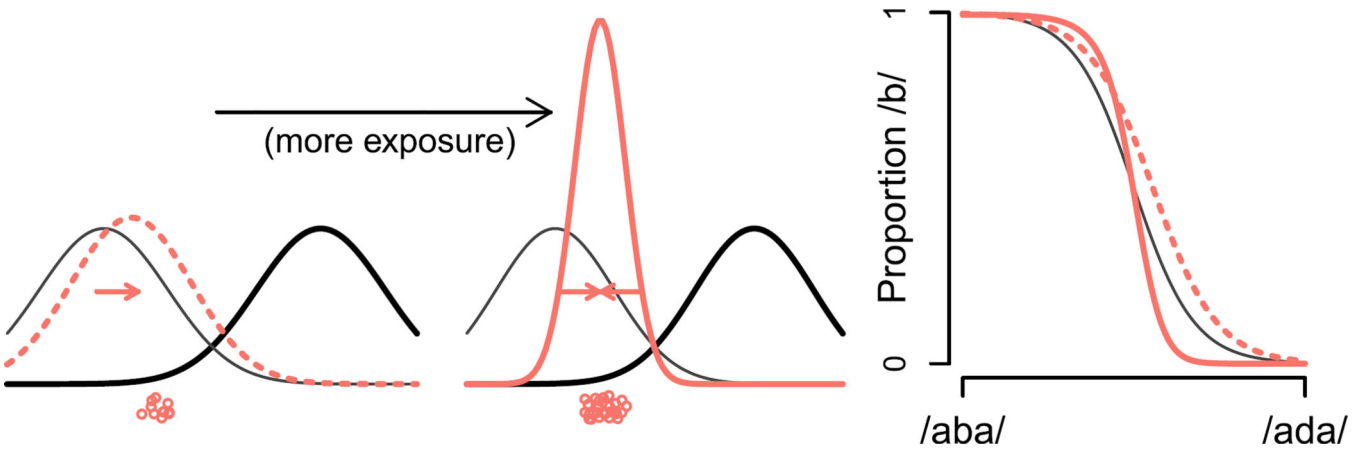


Figure 11.

Schematic illustration of the predicted trade-off between shifts in mean and change in variance. With exposure to tightly clustered (or repeated) stimuli which are perceived as not fully ambiguous, the ideal adapter predicts that the initial shift in mean should lead to a positive aftereffect with small amounts of exposure (left, and dashed line), while the low variance of the repeated adaptor eventually leads to a neutral or even negative aftereffect with prolonged exposure (middle).

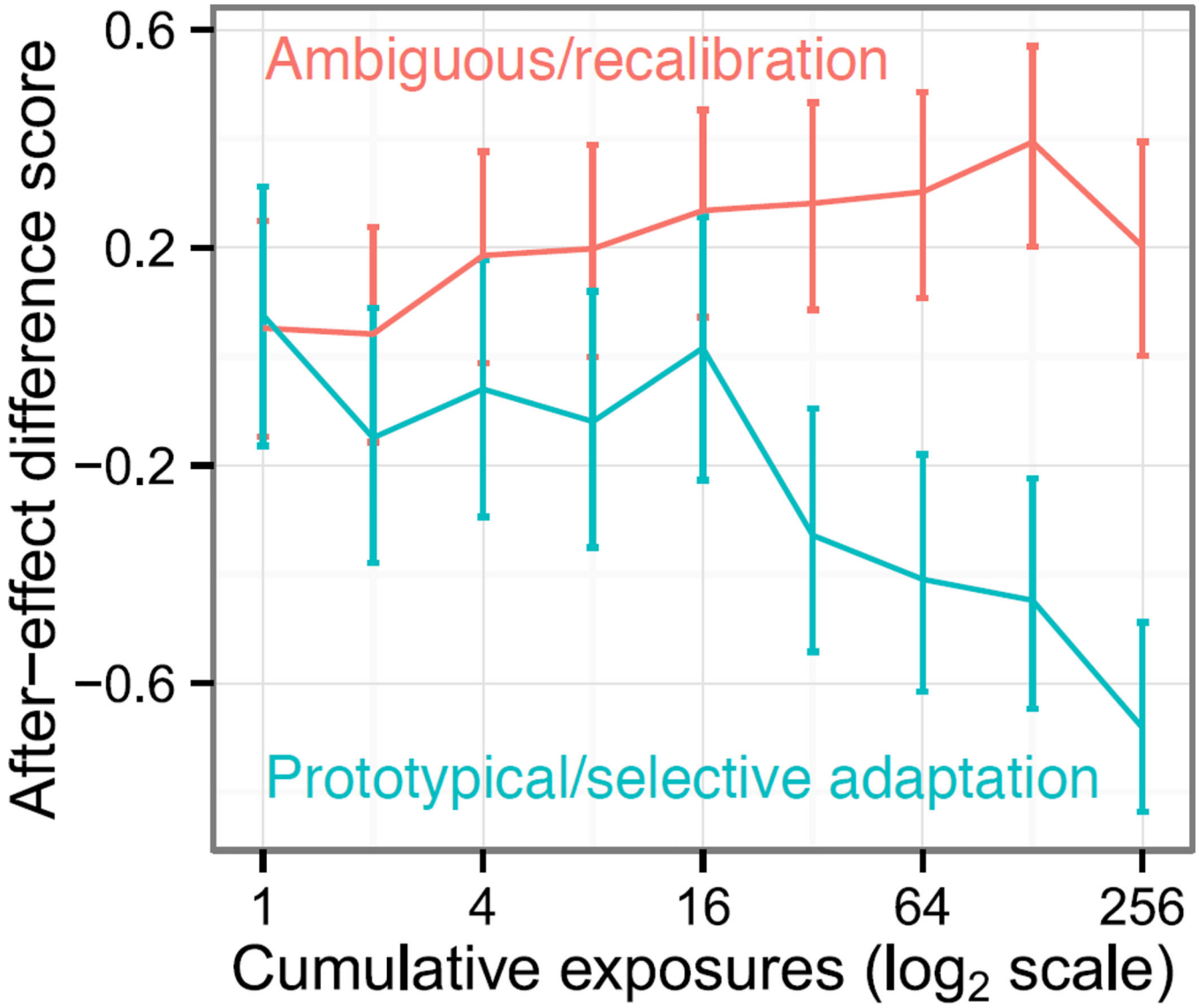


Figure 12.
Results from Vroomen et al. (2007), showing the full 256 exposures. Red/top curve: ambiguous audio-visual adaptor (recalibration). Blue/bottom curve: prototypical audio-visual adaptor (selective adaptation).

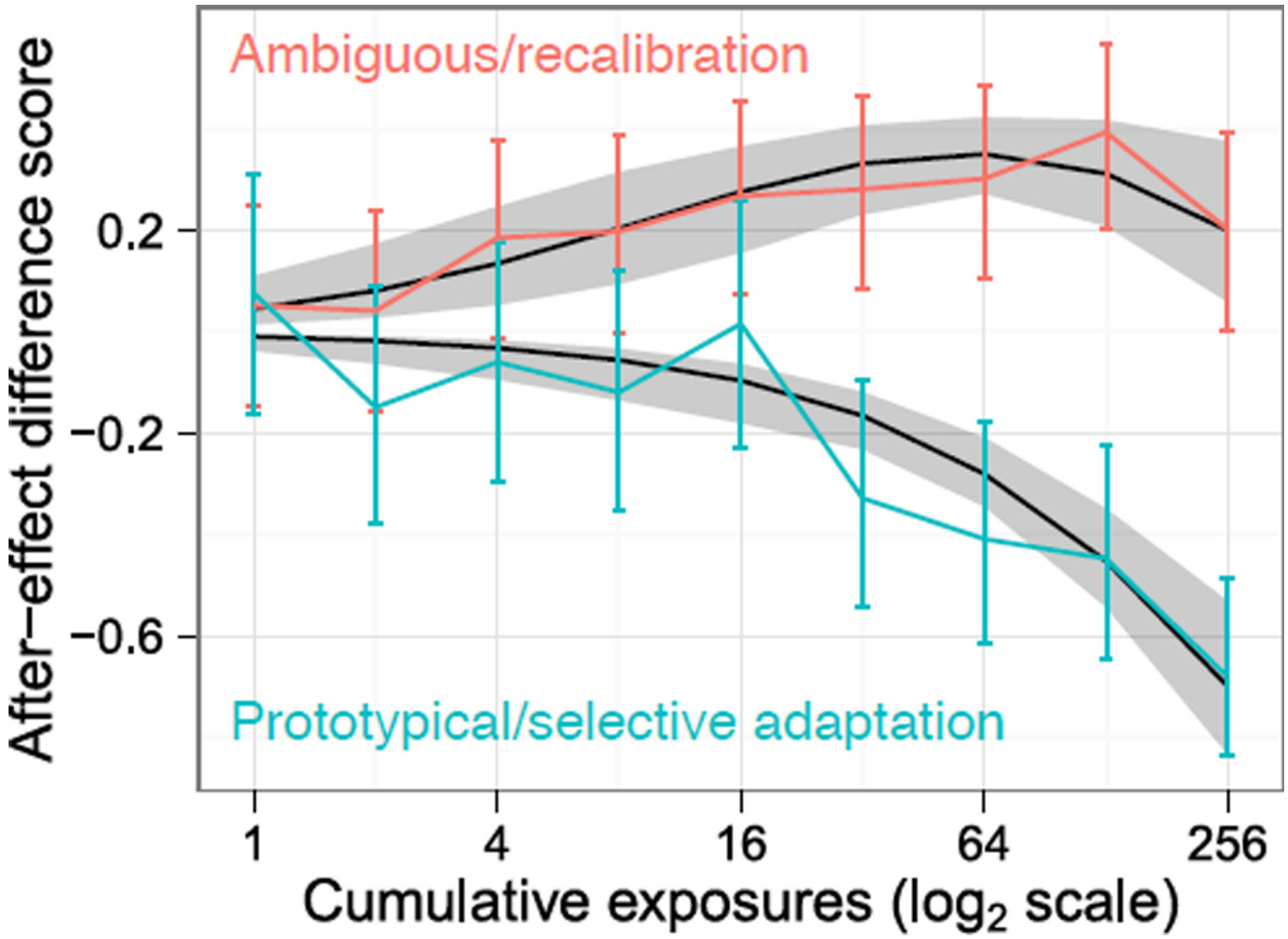


Figure 13.

Results from fitting the belief updating model to all 256 exposures in both conditions from Vroomen et al. (2007) simultaneously. Model predictions correspond to MAP-estimate hyperparameters of $\nu_0 = 100$, $\kappa_0 = 17$, and $w = 0.47$.

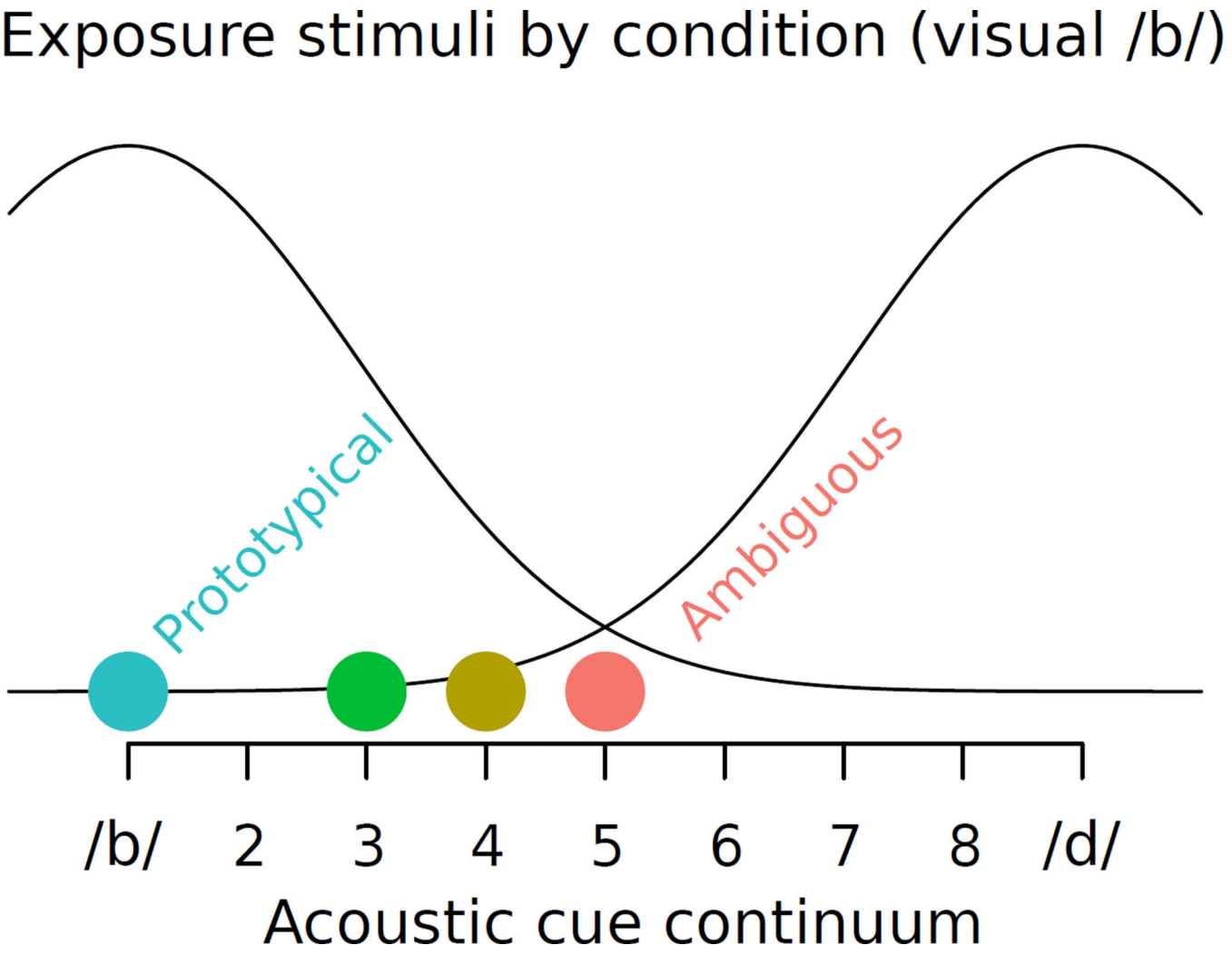


Figure 14.
Construction of stimuli for four conditions with visual /b/ (visual /d/ is analogously the mirror image).

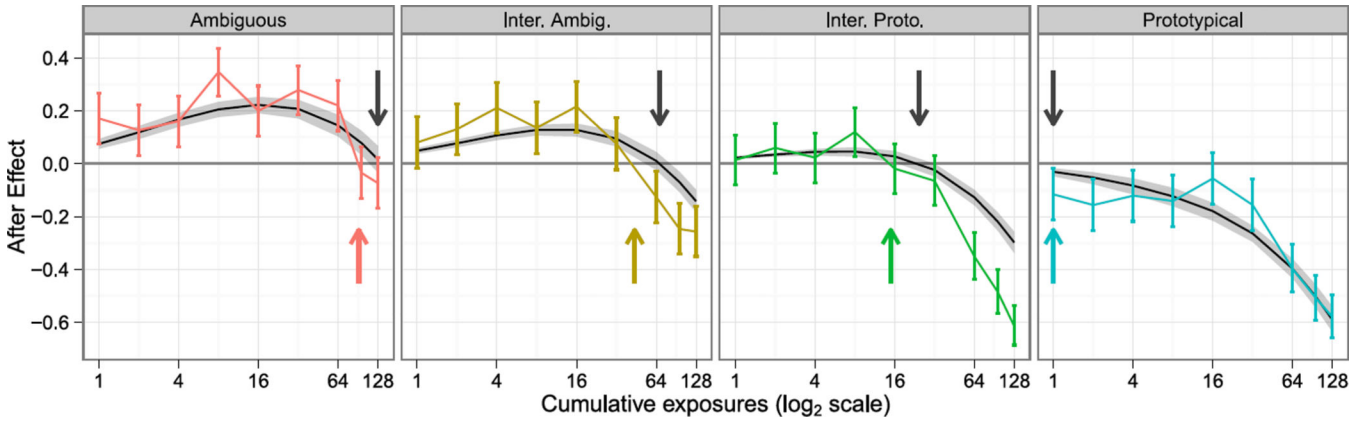


Figure 15.

Results from all four conditions in the first exposure block: ambiguous, intermediate-ambiguous, intermediate-prototypical, and prototypical (colored lines with error bars showing 95% confidence intervals). Model fits (black lines with 95% confidence interval ribbons) are generated based *only* on the ambiguous and prototypical conditions ($r^2 = 0.91$); for the two intermediate conditions the model makes the predictions shown ($r^2 = 0.96$). Colored arrows show the amount of exposure required for behavior to switch from recalibration-like (positive after-effect) to selective adaptation-like (negative after-effect), which decreases as the adaptor stimulus goes from ambiguous to prototypical, as predicted by the model (black arrows; see Figure 16).

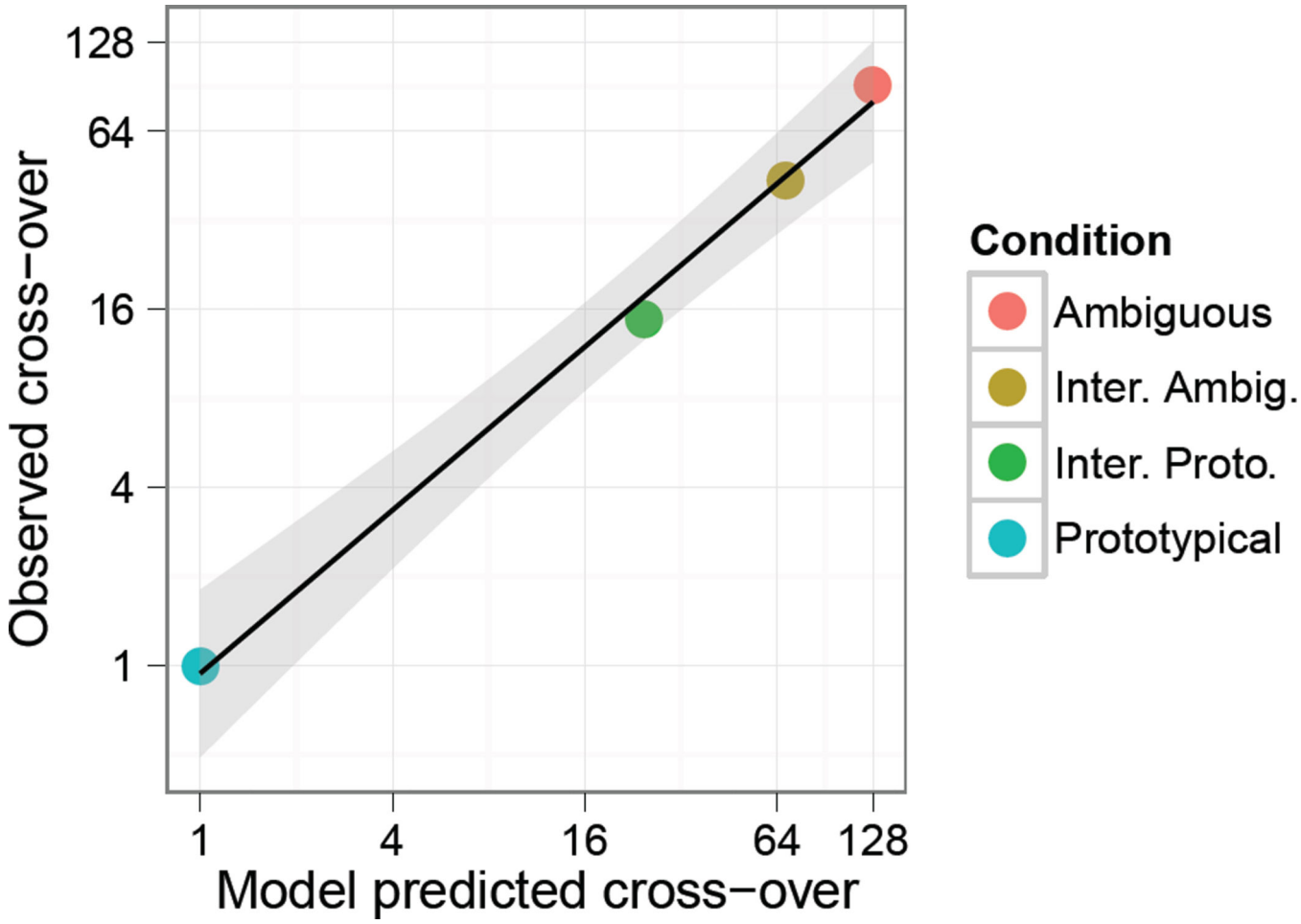


Figure 16.

The belief-updating model predicts when behavior should switch from recalibration-like to selective adaptation-like (see Figure 15). Each dot shows the model predicted cross-over point (*x*-axis) versus the actual observed cross-over point (*y*-axis), with a linear regression fit to these four points showing close agreement.

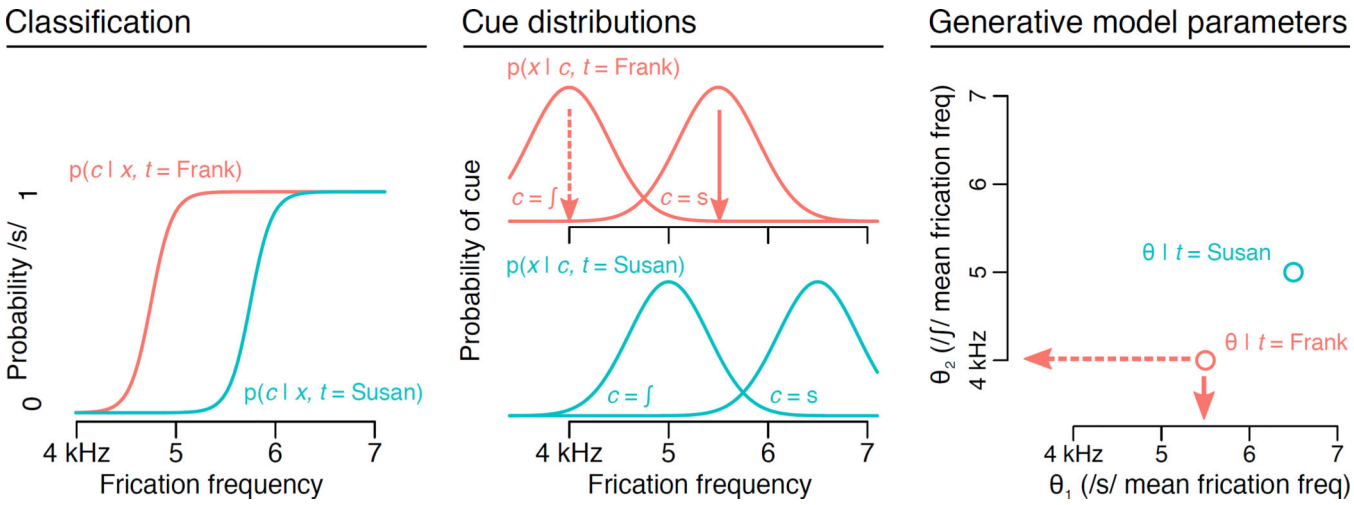


Figure 17.

Talker-specific speech perception can be formalized as inference under uncertainty conditional on talker identity. The probability that a particular cue value *x* (here, frication frequency) was intended to be category *c* (/s/ or /∫/) depends on the talker *t* that produced it, written $p(c \mid x, t)$ (left). This probability is related via Bayes Rule to the talker-specific likelihood, the distribution of cues produced by talker *t* for each category, $p(x \mid c, t)$ (middle). These distributions can be described by the *parameters* of the generative model, θ, such as the talker's mean frication frequency for /s/ and /∫/ as plotted (right). Although we only plot two parameters, many more are required to even approximate the full generative model. Each talker can be thought of as a point in this (very high dimensional) parameter space.

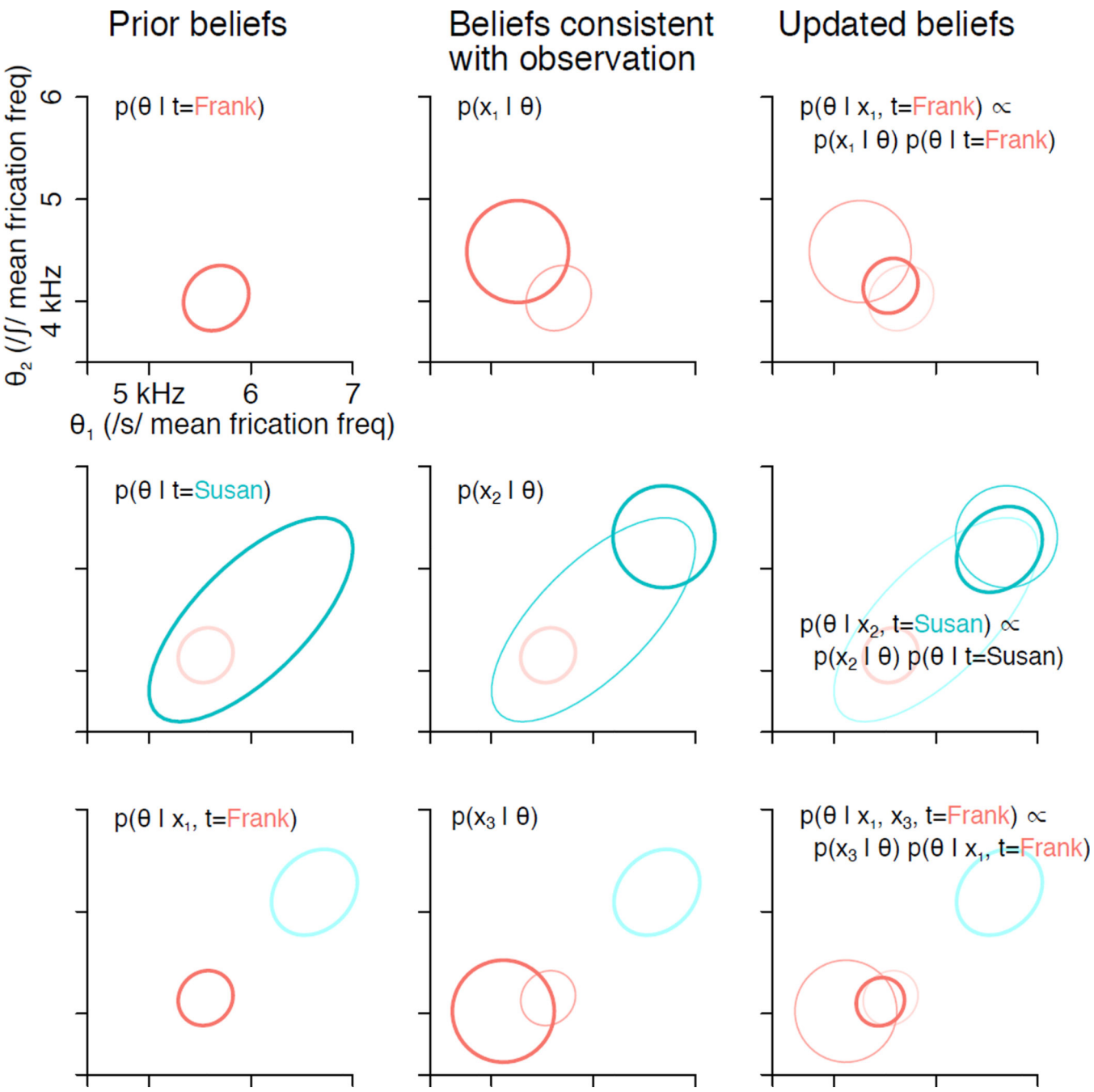


Figure 18.

Adapting to multiple talkers: Updating of talker-specific beliefs allows listeners to continue to learn about individual talkers' generative models by accumulating evidence over multiple encounters. These plots visualize beliefs about generative models as distributions in generative model parameter space (here, showing only mean frication frequencies for /s/ and /ʃ/). Each panel plots equiprobability contours of a distribution (that outline the highest probability region of generative model parameter space), based on prior experience (left column), a single observation (middle), and their combination after belief updating (top). Top row: updating beliefs about a familiar talker ("Frank"), starting from relatively specific

beliefs (left). A single observation x_1 is compatible with a wide range of generative models (middle), but when combined with prior beliefs leads to more specific updated beliefs. Middle row: updating beliefs about a new talker ("Susan"), encountered next, based on very vague prior beliefs and the next observed speech x_2. Bottom row: continuing to update beliefs about Frank—which are not affected by the intervening speech from Susan—after observing another cue value x_3.

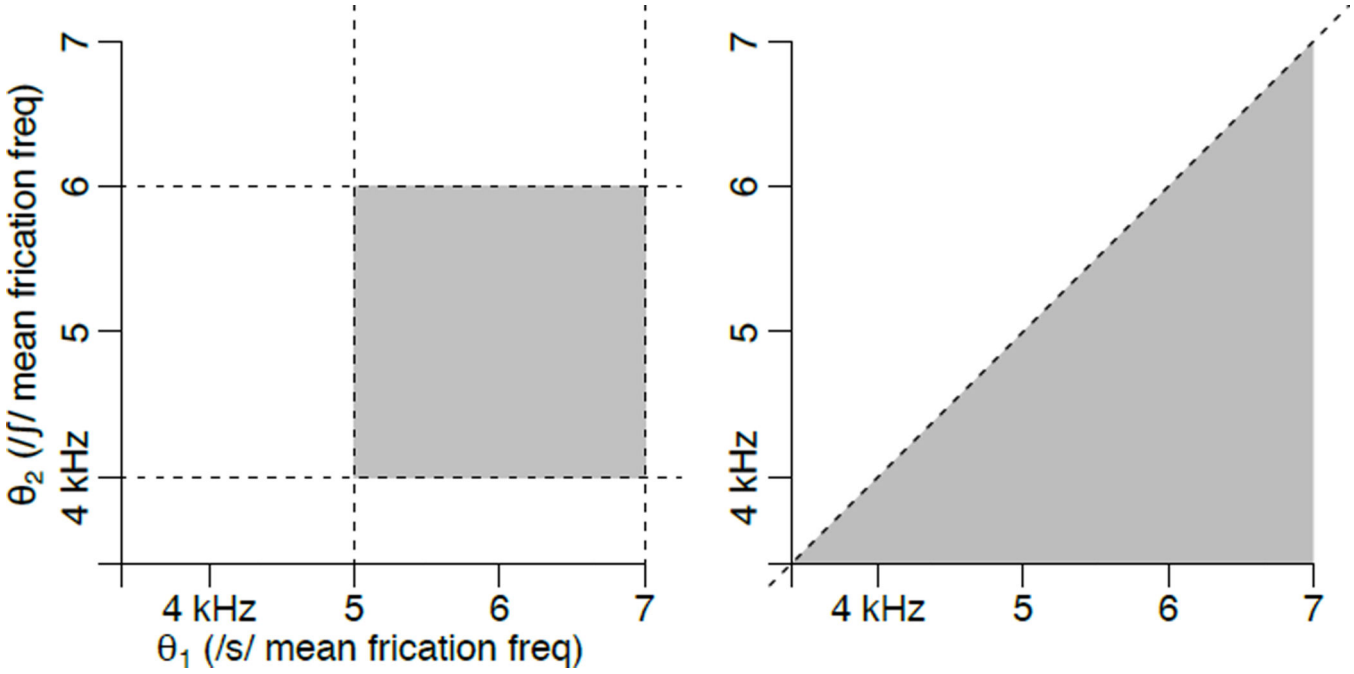


Figure 19.

Two examples of how prior experience with many talkers of a language can constrain the space of generative models that has to be searched during adaptation. Left: an individual talker's mean frication frequency for /s/ generally falls in the 5 kHz to 7 kHz range, while the mean for /ʃ/ is typically in the 4 kHz to 6 kHz range. Combined, these exclude much of the logically possible space of generative models. Right: moreover, the mean for /s/ is generally higher than the mean for /ʃ/, which also excludes a substantial proportion of the possible generative models.

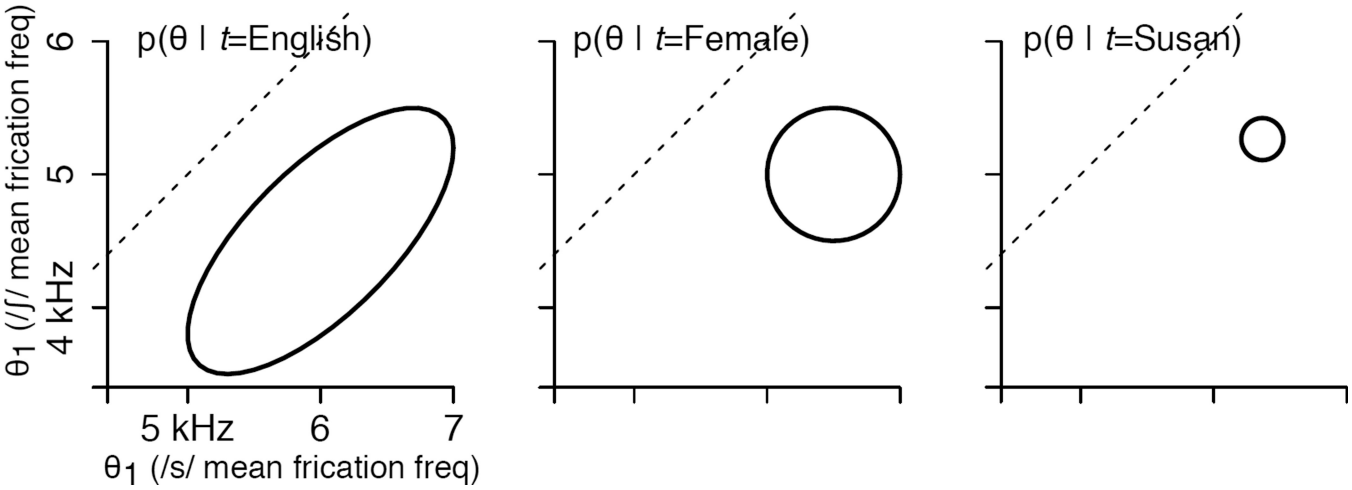


Figure 20.

Prior beliefs at varying levels of specificity, formalized as conditional distributions over generative model parameters (plotted as ellipses of most probability mass). Left: Experience with all talkers of English constrains the listener's prior beliefs about another talker of English, but only very broadly: the mean frication centroid for /s/ tends to be higher than for /ʃ/, and they tend to be positively correlated, but the actual values can range quite a bit. Middle: experience with all female talkers provides more information, because female talkers tend to pronounce /s/ and /ʃ/ with frication frequency means on the high side of the overall range. Thus the distribution of females' generative model parameters for /s/ and /ʃ/ is more concentrated than that of all English speakers. Right: experience with a particular female talker provides even more information, and the corresponding distribution over generative model parameters is even more concentrated