Speech perception as probabilistic inference under uncertainty based on social-indexical knowledge

Kodi Weatherholtz[1],
Maryam Seifeldin[1],
Dave F. Kleinschmidt[1],
Chigusa Kurumada[1],
T. Florian Jaeger[1,2,3]

[1] Department of Brain and Cognitive Sciences, University of Rochester
[2] Department of Computer Science, University of Rochester
[3] Department of Linguistics, University of Rochester

Corresponding Author:
Kodi Weatherholtz, PhD
Department of Brain and Cognitive Sciences
University of Rochester
Meliora Hall, Box 270268
Rochester, NY 14627-0268
Tel: 585-295-0751
email: kweatherholtz@bcs.rochester.edu

Maryam Seifeldin
University of Rochester
Meliora Hall, Box 270268
Rochester, NY 14627-0268
email: mseifeld@ur.rochester.edu

Dave F. Kleinschmidt
University of Rochester
Meliora Hall, Box 270268
Rochester, NY 14627-0268
email: dkleins2@ur.rochester.edu

Chigusa Kurumada
University of Rochester
Meliora Hall, Box 270268
Rochester, NY 14627-0268
email: ckuruma2@ur.rochester.edu

T. Florian Jaeger
University of Rochester
Meliora Hall, Box 270268
Rochester, NY 14627-0268
email: fjaeger@ur.rochester.edu

Word count: 10,184 (excluding abstract, references and appendices)

Abstract

Talkers differ from each other in how they pronounce the same phonetic contrast. In speech perception, such inter-talker variability contributes to the *lack of invariance* problem, creating uncertainty about the mapping between acoustic cues and linguistic representations. However, inter-talker variability is not random: talker-specific cue-to-category mappings are often systematically conditioned on social group membership (including, e.g., a talker's gender and age, but also sociolect and dialect). There is now substantial evidence that listeners take advantage of these statistical contingencies. We provide an introduction to how such sensitivity can be productively understood in terms of ideal observer approaches—mathematical models that describe statistically optimal solutions to inference under uncertainty. We then discuss a challenge that any model of speech perception needs to address: how do listeners know *when* to adapt or, put differently, what variables should cue-to-category mappings be conditioned on? We link this question to the informativity that indexical cues carry with respect to the relevant linguistic distributions, and we demonstrate how the informativity of indexical cues with respect to linguistic distributions can be quantified. If talker identity or social group membership carry important information about the distribution of phonetic categories, ideal observer approaches predict talker- and group-specific expectations. We illustrate how these accounts work against realistic speech data.

Keywords: probabilistic inference under uncertainty, distributional learning, speech perception, socio-indexical knowledge, prediction, generalization, ideal observer models

Acknowledgments

This work was supported by an NIH NRSA Individual Predoctoral Fellowship to DFK (HD082893-01) and an NIHCD R01 (HD075797) to TFJ. The views expressed here are those of the authors and do not necessarily reflect those of the funding agencies.

1. Introduction

The apparent ease and robustness of spoken language understanding belie the considerable computational challenges involved in mapping speech input to linguistic categories. One of the biggest computational challenges stems from the fact that talkers differ from each other in how they pronounce the same phonetic contrast. One talker's realization of /s/ (as in "seat"), for example, might sound like another talker's realization of /ʃ/ (as in "sheet") (Newman et al., 2001). In speech perception, such inter-talker variability contributes to the *lack of invariance* problem, creating uncertainty about the mapping between acoustic information and linguistic representations (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967). A number of proposals for how listeners overcome this problem have been offered (for a recent review, see Weatherholtz & Jaeger, to appear). Although many important questions remain, one important insight that has emerged from this research is that listeners seem to take advantage of statistical contingencies in the speech signal. These contingencies result in part from the fact that inter-talker variability is not random. Rather, inter-talker differences in the cue-to-category mapping are systematically conditioned by a range of factors, such as talker-specific vocal anatomy and articulatory routines (Fitch & Giedd, 1999; Johnson, Ladefoged, & Lindau, 1993) and a talker's social group memberships, including age (Lee, Potamianos, & Narayanan, 1999), gender (Perry, Ohde, & Ashmead, 2001; Peterson & Barney, 1952), and dialect (Labov, Ash, & Boberg, 2006; for recent reviews of talker variability in speech perception, see Foulkes & Hay, 2015; Weatherholtz & Jaeger, to appear).

Listeners seem to draw on these statistical contingencies between linguistic variability on the one hand and talker- and group-specific factors on the other. Upon encountering an unfamiliar talker, for example, the speech perception system seems to adjust the mapping of acoustic cues to speech categories to reflect talker-specific distributional statistics (Bejjanki, Clayards, Knill, & Aslin, 2011; Clayards, Tanenhaus, Aslin, & Jacobs, 2008; Idemaru & Holt, 2011; Kleinschmidt & Jaeger, 2015b; McMurray & Jongman, 2011). Listeners also seem to use top-down knowledge about a talker's social group membership(s) to predict the likelihood with which perceived acoustic cue values map to speech categories (e.g., Hay, Warren & Drager, 2006, Niedzielski, 1999; Strand & Johnson, 1996).

A recent proposal, the ideal adapter (Kleinschmidt & Jaeger, 2015b), provides a unifying computational-level explanation for both the ability to draw on implicit knowledge about talker- and group-specific statistics, as well as the ability to learn novel instances of such statistics. The ideal adapter views speech perception as a problem of inference under uncertainty about not only linguistic categories, but also the appropriate statistical model that underlies the presently observed speech input. Our goal here is to provide an intuitive introduction to this perspective on speech perception and to further illustrate one of its advantages—the ability to make clear quantifiable predictions. Specifically, we have three aims. Our first aim is to introduce a reasoning framework based on probability theory and Bayesian inference that provides a way to think through the computational challenges posed by variability and the brain's solution to it (Section 2). We begin by introducing Bayesian *ideal observer* approaches to speech perception (e.g., Clayards, Tanenhaus, Aslin, & Jacobs, 2008; Feldman, Griffiths, & Morgan, 2009; Norris & McQueen, 2008). Ideal observer models have a long tradition in the cognitive sciences (Anderson, 1990; Marr, 1982; see also Geisler, 2003; Jacobs & Kruschke, 2011; Tenenbaum, Kemp, Griffiths, & Goodman, 2011). The ideal observer approach originates in the idea that the brain makes optimal use of available information (in line with the hypothesis of rational cognition, cf. Anderson, 1990). Ideal observer models specify statistically optimal or "rational" solutions to *inference under uncertainty*, such as recognizing categories from noisy, incomplete, or ambiguous input. For speech perception, for example, ideal observer approaches describe the recognition of phonemes or words based on the observed acoustic cue values and probabilistic knowledge of how acoustic cue values are distributed across linguistic categories.

After introducing the logic of ideal observer models in greater detail below, we then introduce the *ideal adapter*—an extension to ideal observers that defines a rational solution to inter-talker variability (Kleinschmidt & Jaeger, 2015b). The central tenet of the ideal adapter framework is that listeners

represent how variability in the acoustic realization of speech sounds (e.g., the distribution of acoustic cue values associated with /s/ vs. /ʃ/) *covaries* with indexical information (e.g., talker identity, or social group membership, such as gender and age). Knowledge of this *covariance* is sometimes called *socio-indexical linguistic knowledge*. According to the ideal adapter framework, listeners draw on implicit socio-indexical linguistic knowledge, such as talker-specific and group-specific distributional statistics, to infer how observed acoustic information maps to speech categories. Notably, Kleinschmidt and Jaeger (2015b) proposed that listeners continuously update or adapt their beliefs about the covariance between linguistic and indexical information as they experience new input (e.g., additional sound tokens from a familiar talker; or speech from a novel talker). Thus, the ideal adapter framework predicts that category inferences change as the short-term and long-term statistics of the environment change. Specifically, this framework predicts that the speech perception system copes with talker variability by (i) *learning* talker-specific and group-specific distributional statistics, and (ii) *storing* this information in memory in order to facilitate subsequent recognition of familiar forms and generalization to novel similar forms.

Our second aim is to relate this reasoning framework to evidence regarding the influence of linguistic variability on speech perception and the ability of listeners to adapt to unfamiliar or otherwise atypical variability (Section 3). Specifically, we assess to what extent the existing speech perception literature supports the prediction that listeners cope with the lack of invariance in speech by learning and storing conditional distributional statistics that are informative about the underlying pattern of variability in the input (e.g., talker- and group-specific distributional statistics). While we focus primarily on talker variability (idiolect, dialect, sociolect), we also discuss briefly how the framework outlined below extends to account for other types of variability, such as within-talker stylistic variability.

Our third aim is to extend this framework to address the specific question of *when* the speech perception system should adapt to variability and when the speech perception system should fail to do so (Section 4). Given that cognitive resources are limited, humans presumably cannot attend to and store information about all factors that contribute to variability in speech. Hence, the brain should have a principled means of determining which factors are informative with respect to linguistic cue and category distributions and, hence, which factors warrant attention. We illustrate how the ideal adapter framework provides a solution to this issue. Specifically, we present a series of novel simulations against real world speech production data that quantify the informativity of conditioning factors at multiple levels: at the level of cue distributions, and the influence of these cue distributions on ideal categorization performance.

Several of the predictions of the ideal adapter outlined above—such as the learning and storage of talker- and group-specific statistics—though not necessarily put in these terms, are shared with a number of other approaches to speech perception. This includes episodic accounts (Goldinger, 1997), exemplar-based accounts (Johnson, 1997; Pierrehumbert, 2006, 2002), and certain pre-categorical normalization accounts (McMurray & Jongman, 2011; Cole, Linebaugh, Munson & McMurray, 2010; Huang & Holt, 2009; Holt, 2005). Here we focus on the ideal adapter. This approach has only recently entered research on the lack of invariance problem. We thus seek to provide the reader with general intuitions about how key concepts—such as *distributional learning*, *belief updating*, *inference under uncertainty* and *varying (non-stationary) statistics*—relate to central questions about variability and language processing, independent of the (potentially many and diverse) cognitive mechanisms or neural algorithms that might implement the relevant computations. In doing so, we hope to illustrate what ideal observer/adapter models can offer to researchers in speech perception. This includes the ability to derive quantitative predictions based on prior principles of learning that should be observed independent of the specific mechanistic framework in which these principles are implemented. For example, since ideal observer models are statistically optimal, they provide an upper bound against which to evaluate human performance, thereby facilitating interpretation of actually observed human performance. Ideal observers can provide an informative baseline for the development and comparison of different mechanistic models, too. Further, the ideal adapter can help to explain *why* a specific model of speech perception and adaptation does or does not explain human behavior—e.g., whether it is because the mechanism resembles that employed by humans or because the model is one of many algorithmic variants that

achieves the goal of rational learning.[1] Specifically, the ideal adapter highlights the primary challenges the speech perception system needs to overcome. Finally, the ideal adapter framework for speech perception is computationally related to ideal observer-style models at higher levels of language processing (Kuperberg & Jaeger, 2016; Norris, McQueen, & Cutler, 2015; Gibson et al., 2013; Bicknell et al., 2012). One advantage of this connection is that it facilitates the formulation of testable predictions about how inference under uncertainty during speech perception affects higher level processing, such as message comprehension. For all of these reasons, we consider ideal observers and adaptors a valuable theoretical tool in advancing our understanding of speech perception.

2. Speech perception: probabilistic inference under uncertainty based on social-indexical knowledge

We begin by presenting a reasoning framework that characterizes the cue-to-category mapping process in speech perception as Bayesian inference. We first introduce the notion of *inference under uncertainty*. We explain how this notion relates to speech perception by presenting a Bayesian ideal observer account of how listeners infer phonological categories from acoustic information under the simplifying assumption of *stationarity*: i.e., that the distribution of acoustic cue values associated with phonological categories is the same across talkers. We then present the ideal adapter framework (Kleinschmidt & Jaeger, 2015b), which provides an optimal solution for the cue-to-category mapping given that variability in speech is *subjectively non-stationary* (i.e. from the perspective of the listener, the distribution of acoustic cue values associated with phonological categories varies over time).

2.1. Inference under uncertainty

Uncertainty arises in language processing due, e.g., to inter-talker or intra-talker variability in how linguistic categories are produced (e.g., phonetic variability). Thus, to process information efficiently and to make *rational* decisions, humans must be able to cope with uncertainty. Probability theory provides the mathematical tools for representing and manipulating uncertainty.

Assume a listener's current goal is to correctly categorize a perceived sound: i.e., to map the acoustic input to the phoneme intended by the talker. Because there is no one-to-one relationship between acoustic cues and phonological categories (the *lack of invariance* problem), the cue-to-category mapping process is inherently characterized by uncertainty. A Bayesian model of categorization copes with this uncertainty by combining two sources of information—which are called the *likelihood* and the *prior*—to infer the most probable category, given the physical stimulus. The *likelihood* is the probability of observing the perceived stimulus (e.g., the particular combination of perceived acoustic cue values) based on different hypotheses about how those data were *generated*. For categorization problems these hypotheses often correspond to categories (e.g., the hypothesis that the observed percept was generated by the category /s/; or more intuitively, that the percept was *generated by a talker who intended* the category /s/). The *prior* is the overall probability of observing the hypothesized category, based on the perceiver's past experience: e.g., how likely a listener is to observe a given phoneme, independent of the physical properties with which that phoneme was realized. Together, the likelihood and prior (and a normalization term) determine the *posterior probability* of a phoneme given a percept. It is this posterior probability that determines whether an ideal observer recognizes a percept as one phoneme or another.[2]

[1] Ideal observer models are a computational level framework and—at least in their purest form—make unrealistic assumptions, such as unlimited attentional and memory resources, (for further discussion, see also Chater & Oaksford 1998; Griffith, Chater, Kemp, Perfors, & Tennenbaum, 2010; Pouget & Knill, 2004; Simon, 1956,

[2] Specifically, if the costs of miscategorization are uniform across phonetic categories, categorization performance is maximized on average by choosing the category with the highest posterior probability. For cases in which the cost of misclassification is not uniform, the *utility* of categorization is maximized for minimizing the expected cost, which is function of both the posterior and the costs of different misclassifications (for discussion, see Kuperberg & Jaeger, in press, and references therein).

The joint impact of the likelihood and prior on the posterior probability is described by Bayes' Rule, which follows from the basic axioms of probability theory:

$$(1)\quad p(C = c_i \mid \text{percept}) = \frac{p(\text{percept} \mid c_i) * p(c_i)}{\sum_{j=1}^{n} p(\text{percept} \mid c_j) * p(c_j)}$$

Equation 1 indicates the posterior probability of each phonemic category $C = c_i$ given a specific percept, p($C = c_i$ | percept). The numerator on the right-hand side of the equation is the likelihood of the percept under the category c_i (i.e., the probability of observing that percept given the hypothesis that the intended category was c_i) multiplied by the prior probability of the category c_i, p(c_i). The denominator is a normalization constant that ensures that the sum of all posterior probabilities p(C = c_i | percept) is one. For simplicity, the normalization term is often omitted, and the relationship between the posterior probability on the one hand and the likelihood and prior on the other is expressed proportionally:

$$(2)\quad p(\mathrm{C} = \mathrm{c}_i \mid \text{percept}) \propto p(\text{percept} \mid \mathrm{c}_i) * p(\mathrm{C}_i)$$

Figure 1 illustrates a hypothetical example of optimal categorization using Bayesian inference. For this example, we assume the listener's goal is to determine whether a physical stimulus is an instance of /b/ or /p/ and that the prior probability of observing these two categories is equal. For simplicity's sake we consider only the effect of *one* acoustic cue, voice onset time (VOT). VOTs are the primary cue to voicing contrasts in English onset plosives (e.g., the contrast between /bin/ and /pin/).[3] The top panel shows the distribution of VOTs associated with each phoneme based on the listener's prior experience: we can think of this as an approximation of the distribution of cue values across all previously observed exemplars of each category, independent of the talker. The bottom panel shows the Bayes optimal classification function for identifying a physical stimulus as /b/, given the observed VOT value, under the assumptions of uniform prior probabilities and uniform costs of miscategorization for /b/ and /p/, cf. footnote 2). Under the assumption that the prior probabilities of /b/ and /p/ are equal, the *posterior probability* of /b/ given an observed VOT value is proportional to the *likelihood* of observing that VOT value given the known distribution of cue values associated with /b/. As shown in Figure 1, the posterior probability of /b/ is 0.5 at the point of maximal overlap between the two categories, and the posterior probability of /b/ increases (or decreases) to the left (or right) of the point of maximal ambiguity.

[3] Additional cues contribute to the perception of voicing in English onset plosives (Toscano & McMurray, 2010; Stevens & Klatt, 1974). The arguments we present here readily extend to such cases, in which categories form multi-dimensional cue distributions. Indeed, we present such examples below.

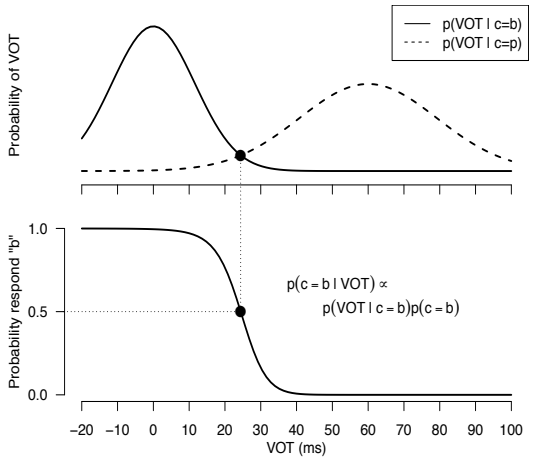

Figure 1: Hypothetical distribution of voice onset times for tokens of /b/ and /p/ (top) and the ideal (statistically optimal) classification function for mapping VOT to /b/. For simplifying assumptions, see text.

Bayesian ideal observer models are increasingly influential in research on language processing, including speech perception (Bejjanki et al., 2011; Clayards et al., 2008; Feldman et al., 2009; Sonderegger & Yu, 2010) and spoken word recognition (Norris, 2006; Norris, McQueen, & Cutler, 2015; Norris & McQueen, 2008). Related Bayesian models have been proposed for parsing (e.g., Bicknell et al., 2012; Gibson et al., 2013; Jurafsky, 1996; Levy, 2008, 2011) and, more generally, as a framework for language understanding (for review, see Kuperberg & Jaeger, 2016).

Like early ideal observer models in many other cognitive domains, these models of language processing make the simplifying assumption of stationarity. That is, these models assume that the statistical distributions of linguistic categories do not differ across contexts. This assumption is violated by the presence of linguistic variability that is conditioned on *contextual* information. By context, we mean both *linguistic context* (e.g., variability due to phonotactics and co-articulation), as well as *extra-linguistic context*, such as *indexical or social context* (e.g., variability due to talker-specific, group-specific or stylistic differences). At least for speech perception, variability due to socio-indexical context arguably constitutes a larger computational problem than variability due to linguistic context: the number of relevant phonological contexts that the realization of phonetic contrasts is conditioned on is comparatively small, whereas we continue to encounter new talkers and new types of talkers (groups) throughout our life. We turn now to a discussion of how socio-indexical knowledge affects speech perception within the ideal observer tradition.

2.2. Socio-indexical linguistic knowledge constrains the likelihood and the prior

Under the simplifying assumption of stationarity, the same statistical distributions would be used to derive the likelihood and prior regardless of the talker who produced the utterance. We know empirically, however, that the same acoustic cue values map to different categories with different probabilities depending on the talker. One talker's realization of /s/, for example, can be physically very similar to another talker's realization of /ʃ/, due to such factors as talker-specific idiosyncrasies and gender-based vocal tract differences (Newman et al., 2001). In other words, the *likelihood* that a phonological category is realized with particular acoustic cue values is *non-stationary* from the listener's perspective. The *prior* probability of phonological categories is, likewise, non-stationary. For example, certain talkers or groups of talkers might be more or less likely to use certain phonemes (e.g., depending on the phonological composition of the words they use most frequently in their daily life or vocation). Here we show how conditioning probability estimates on socio-indexical knowledge, such as talker-specific and group-specific distributional statistics, affects category inferences in an ideal observer framework.

As discussed above, variability abounds in language, but this variability is not random. Rather variability in language is at least partially systematic. Consider variability due to vocal anatomical differences across talkers. Differences in the size and shape of the vocal tract lead to considerable inter-

talker variability in the acoustic realization of speech sounds (Hillenbrand et al., 1995, Fitch & Giedd, 1999, Peterson & Barney, 1952). At the same time, however, vocal tract characteristics are a source of within-talker stability. For example, the spectral distribution of energy that characterizes a talker's realization of, say, the vowel /i/ will tend to be similar across tokens of this category because the energy is resonating through a tube with a relatively fixed size and shape.[4] Further, group-level factors such as a talker's sex, regional background and social group memberships, lend higher-level systematicity: talkers who share a dialect or sociolect, for example, by definition produce similar patterns of variation (e.g., whether the vowels in the words *pin* and *pen* are merged or distinct).

The fact that variability is *systematically* conditioned by talker-specific and group-specific factors means that variability contains *information* with respect to those factors. This information can be helpful for language processing, once it is learned. Specifically, once listeners learn how variability along a particular linguistic dimension covaries with socio-indexical factors, listeners can use knowledge of this covariance to make more precise inferences.[5] The basic intuition is that instead of estimating the likelihood and the prior based on all category exemplars that a listener has ever experienced, it can be advantageous to estimate the likelihood and prior talker-specifically or group-specifically. E.g., how likely is it that the observed acoustic cue values were generated by a talker who intended to produce the vowel /i/, given that the listener knows the talker is female and that the listener has knowledge of gender-related vowel variability? We first illustrate this idea for a simple case, using hypothetical data. Then we turn to a more complex case, the English vowel space, and show how an ideal observer framework can help us quantify the (predicted) advantage of learning and storing talker- or group-specific statistics.

2.2.1. A simple hypothetical example: Talker-specific variability in voice onset time

Figure 2 illustrates the change in statistically optimal categorization performance depending on whether talker-specific differences (in this case, in the likelihood distribution) are considered or not. The colored lines in the top panel of Figure 2 show hypothetical VOT distributions for /b/ and /p/ from different talkers. The blank lines show the corresponding *average* or *marginal* distribution across talkers, obtained under the assumptions that (i) each talker contributed the same amount of data, and (ii) the prior probability of /b/ and /p/ does not differ across talkers (i.e., assuming that talkers do not differ in the frequency with which they produce these two phonemes). The thick black classification function in the bottom panel shows the ideal solution for mapping VOT values to /b/ under the assumption of stationarity—i.e., the ideal solution when inter-talker variability is ignored and the marginal distribution across talkers is used to derive the mapping function. The thin colored classification functions show the ideal solution for VOT variability when perfect knowledge of the talker-specific statistics (and perfect certainty about talker identity) is assumed. Note that the population-level classification function provides a poor fit to the talker-specific data.

[4] We do not mean to suggest that the size and shape of a talker's vocal tract are fixed anatomical characteristics. These properties change over time (e.g., due to the lowering of the larynx when males go through puberty, which lengthens the vocal tract overall), and talker's can strategically manipulate these properties (e.g., by pursing their lips to temporarily lengthen their vocal tract). Our point is that when comparing across utterances, the degree of vocal tract-related variability is likely to be dramatically smaller when those utterances are produced by the same talker than by different talkers. For empirical validation of this idea, see also Section 2.2.2.

[5] This insight was acknowledged early on in research on speech perception (e.g., Liberman & Mattingly, 1985; see also Pisoni, 1997). At least implicitly, this insight has arguably also been at the heart of certain normalization accounts (e.g., Ladefoged & Broadbent, 1957; Holt, 2005: and, in particular, Cole, Linebaugh, Munson, & McMurray, 2010; McMurray & Jongman, 2011; for a recent review, see Weatherholtz & Jaeger, to appear). The ideal adapter we discuss below derives this insight from prior principles and integrates it into the inference process (Kleinschmidt & Jaeger, 2015b).

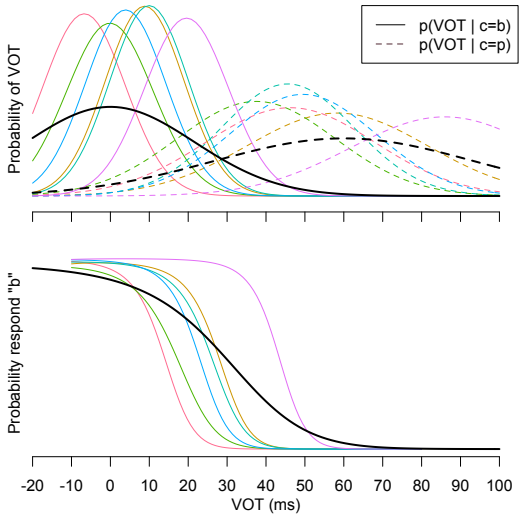

Figure 2: Top panel shows hypothetical distributions of VOT for /b/ (solid lines) and /p/ (dashed lines). Thin colored lines show talker-specific distributions, and thick solid lines show the corresponding *average* or *marginal* distributions across talkers. Bottom panel shows the ideal (statistically optimal) classification function for mapping VOT values to the category /b/ based on talker-specific VOT distributions (colored lines) or the marginal population-level distributions (solid black line). Note that the population-level classification function provides a poor fit to the talker-specific data.

The example in Figure 2 is based on hypothetical data that varies along a single dimension. However, speech is high dimensional, and speech categories are complex constellations of acoustic-phonetic features that can vary simultaneously along multiple dimensions. We next illustrate how socio-indexical linguistic knowledge can affect inferences about speech categories when category-relevant variability spans multiple acoustic dimensions. We focus on vowel categories, which in English are cued primarily by formant frequencies (resonance frequencies of the vocal tract).

2.2.2. Quantifying the advantage of talker-specific vowel knowledge

We use the vowel corpus collected by Heald and Nusbaum (2015) to illustrate how knowledge of talker-specific distributional statistics affects vowel categorization. The corpus consists of recordings from eight talkers producing seven vowels (/i, ɪ, ɛ, æ, u, ʌ, ɑ/) on multiple days and at various times throughout each day. Figure 3 illustrates the vowel spaces of the eight talkers along the first and second formant (F_1 and F_2).[6] Note that the talker-specific vowel distributions (colored contours) differ considerably across talkers; as a result, the marginal distributions (grey contours) provide a poor fit to the talker-specific data. The difference between the talker-specific and marginal distributions can be measured by the *Kullback-Leibler (KL) divergence*, an information-theoretic measure of the information lost when using one distribution to estimate another. Figure 4 shows that there is significant information loss when using the marginal (population-level) distributional statistics in Heald and Nusbaum's corpus to estimate the talker-specific vowel distributions: the average information loss is about 2.3 bits per vowel (see Appendix A for details of this analysis, including the formulae for calculating KL divergence in the general case and in the specific case of multivariate Gaussians). Out of context this number is hard to interpret and we return to this point in Section 4, where we compare the informativity of different socio-indexical variables. Here we simply note that the information loss is significantly larger than zero (as indicated by the bootstrapped confidence intervals in Figure 4).

[6] Figure 5 was derived by sampling 100 vowel tokens of each vowel category from each talker based on the talker-specific distributional statistics for F_1 and F_2 (means, variances and covariance), under the assumption that vowels form multivariate normal distributions in F_1xF_2 space. The talker-specific distributional statistics were generously provided by Shannon Heald. Human Subject Protocols did not allow us access to the raw data.

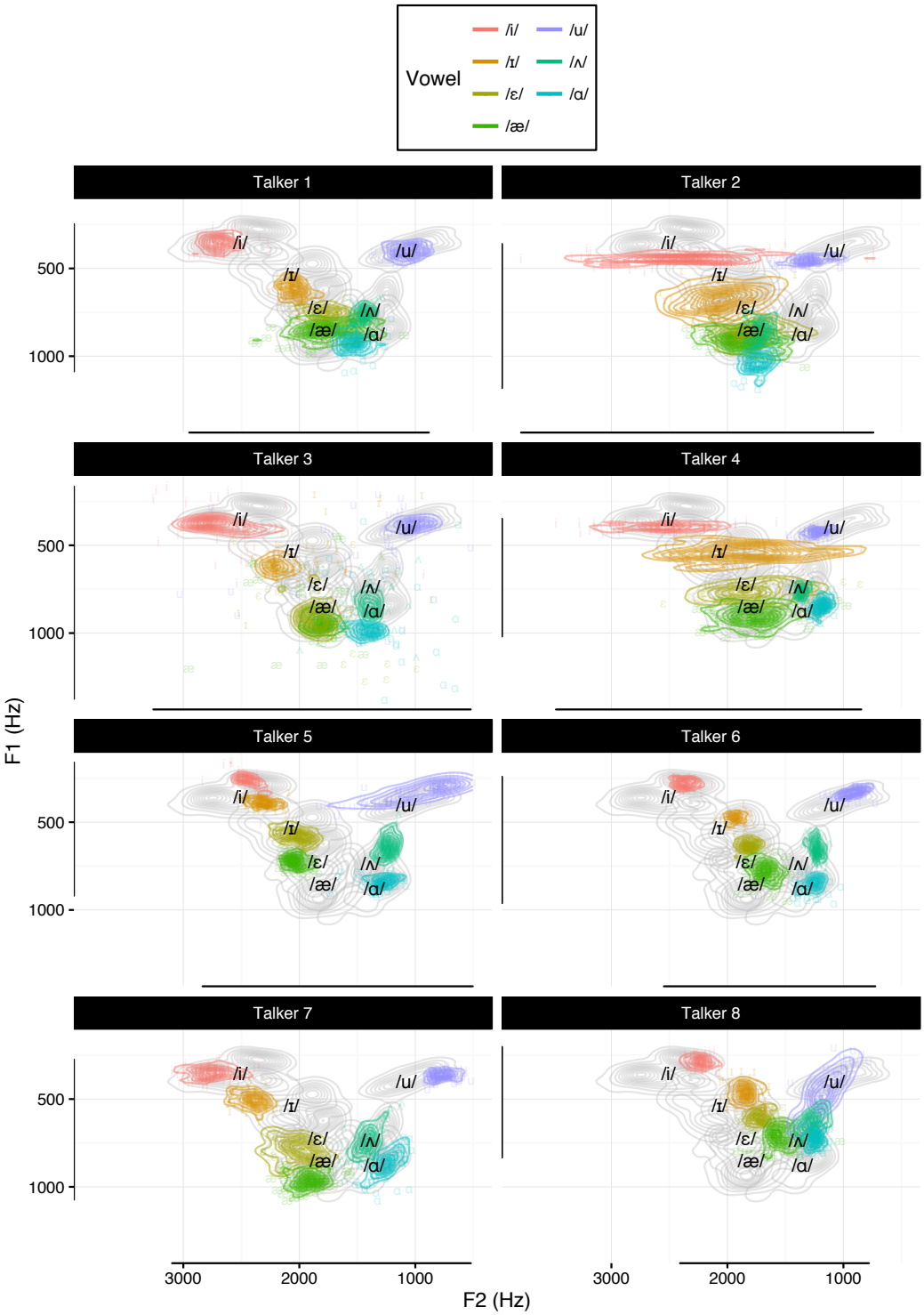

Figure 3: Simulated vowel space for the eight talkers in Heald and Nusbaum's (2015) corpus, assuming a multivariate normal F_1xF_2 distribution for each vowel. The colored contours in each panel indicate talker-specific vowel distributions. The grey contours indicate the marginal vowel distributions across talkers (i.e., the grey contours are the same in each panel). The black vowel labels indicate the mean F_1 and F_2 for each vowel under the marginal distribution. The eight talkers have markedly different vowel spaces, despite speaking with the same (Inland North) dialect of American English.

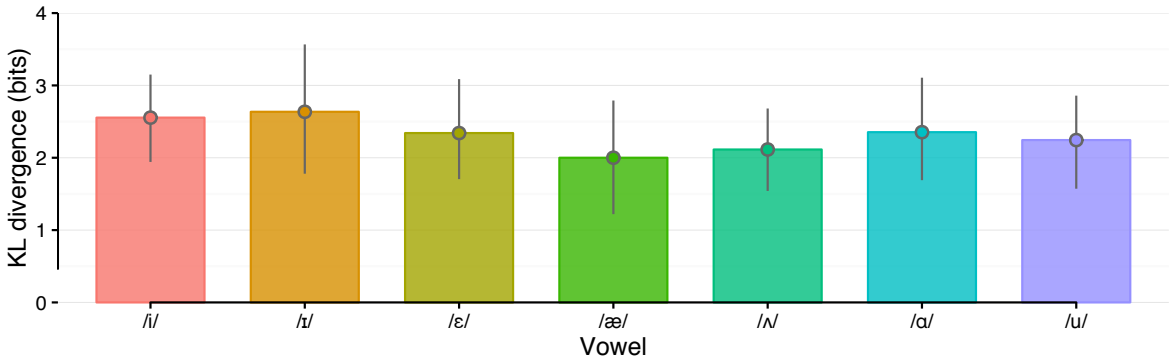

Figure 4: Average information loss (KL divergence) when using the marginal F_1xF_2 distributional statistics from Heald and Nusbaum's (2015) corpus to estimate the talker-specific distributions.

Critically, the information loss illustrated in Figure 4 has consequences for everyday speech perception. One of the advantages of the ideal observer approach is that we can straightforwardly estimate these consequences: we repeatedly sampled vowel tokens from the talker-specific distributions and then asked how likely each vowel token was to be correctly recognized under an ideal observer that had access to talker-specific statistics, compared to an ideal observer that only had access to the marginal (average) distribution across talkers. Note that each vowel token is categorized as the vowel category with the highest posterior probability given the observed acoustic cues (i.e., the *criterion* rule; the results are robust to different choice rules, though; for details, see Appendix B). Figure 5 shows the results of this comparison: as expected given that talker identity is informative about the distribution of vowel tokens in F_1xF_2 space, the average probability of correctly recognizing vowel tokens was significantly higher when the cue-to-category mapping was conditioned on talker. (Note that this increase in recognition accuracy is particularly large for the vowels /ɛ/ and /æ/, which are among the most confusable vowels in American English; Hillenbrand et al., 1995).

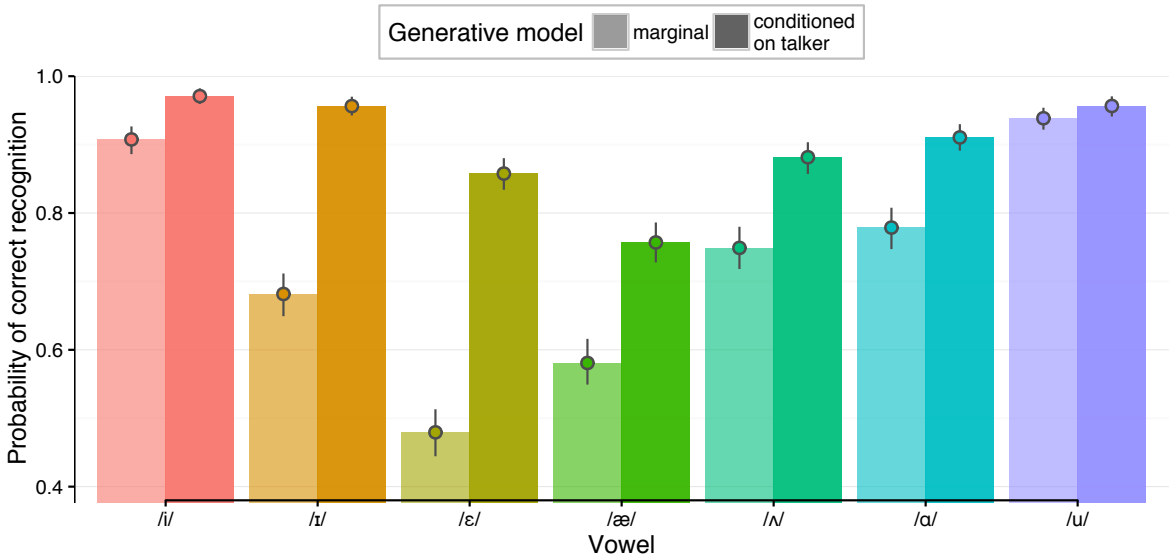

Figure 5: Probability of correctly recognizing vowel tokens under an ideal observer model with perfect knowledge of the talker-specific F_1xF_2 distributional statistics in Heald and Nusbaum's (2015) vowel corpus (dark bars) versus an ideal observer with knowledge of only the marginal distributional statistics (light bars).

Next, we illustrate how the same approach can be applied to estimate the consequences of ignoring talker groups that are predictive of category and cue distributions.

2.2.3. Quantifying the advantage of group-specific vowel knowledge

To illustrate how knowledge of group-specific distributional statistics can affect category inferences, we use male and female vowel productions from the Peterson and Barney (1952) vowel corpus. This corpus comprises tokens of 10 vowels produced by 76 talkers from various American English dialect regions. Figure 6 shows the gender difference for a specific category contrast—/ɛ/-/æ/—in F_1xF_2 space: the left panel shows the marginal distribution of vowel tokens across talkers (under the assumption of normality), and the right panel shows the corresponding distributions for males and females separately. Note that the marginal (population-level) probability distributions for these two vowel categories are wide and highly overlapping (left panel). By contrast, when the probability distributions are conditioned on the gender of the talker (right panel), the distributions are considerably more "peaked": that is, the within-category variance is smaller. As a consequence of this peakedness, the classification function of an ideal observer with knowledge of group-specific distributional statistics will be more steep or, put differently, listeners would have less uncertainty about the mapping of percepts to vowel categories.

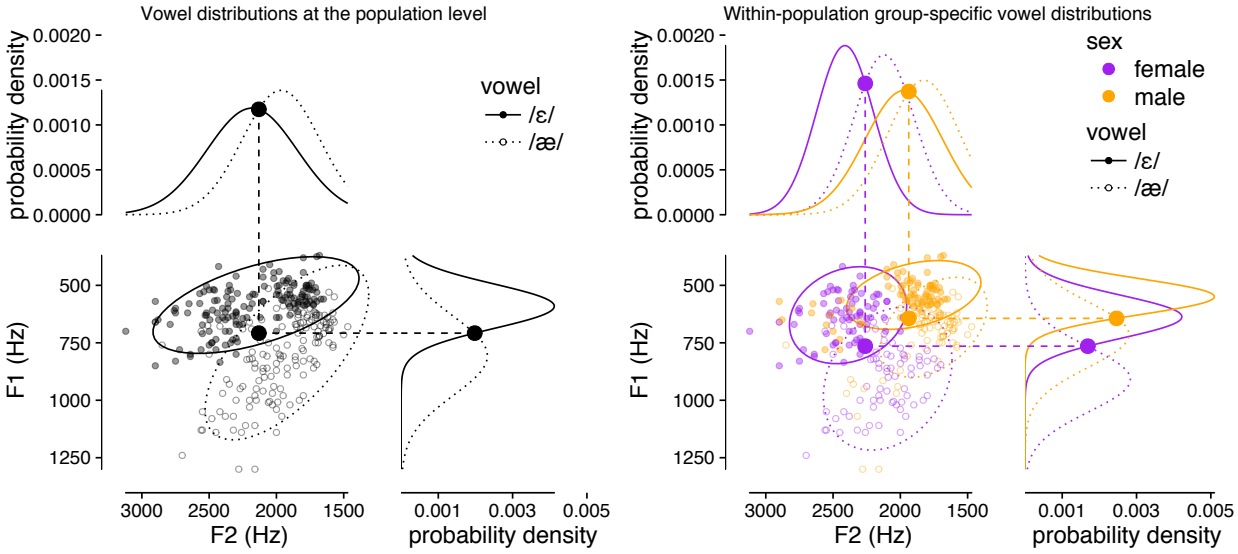

Figure 6: Tokens of /ɛ/ and /æ/ from male and female talkers in the Peterson & Barney (1952) vowel corpus (accessed from the `phonTools` package in R). The left panel shows the marginal distribution of F_1 and F_2 across talkers (under the assumption of normality). The right panel shows the corresponding probability distributions conditioned on gender. The large points indicate the point of maximal ambiguity between /ɛ/ and /æ/ under the marginal and gender-specific distributions. Note that the "peakedness" of the probability distributions is a function of the within-category variance. Since the by-gender distributions are more peaked than the marginal distributions, an ideal observer would have less uncertainty about the cue-to-category mapping when conditioning inferences on the gender-specific rather than marginal statistics.

Figure 7 shows the average probability of correctly mapping each vowel token in the Peterson and Barney (1952) corpus to the category intended by the talker under an ideal observer model with perfect knowledge of the gender-specific F_1xF_2 distributional statistics (dark bars) and under an ideal observer model with knowledge of only the marginal F_1xF_2 distributional statistics (light bars). Indeed, as suggested by the difference in peakedness of the F_1xF_2 distributions in Figure 6, the average probability of correctly recognizing tokens of /ɛ/ and /æ/ is significantly higher when the cue-to-category mapping is conditional on the talker's gender. Note that for all 10 vowels, category recognition accuracy is higher when conditioning on talker gender than when ignoring talker gender; however, the magnitude of this accuracy increase varies considerably across vowels and for some vowels appears to be negligible. This variability by vowel suggests that gender is informative about the mapping of acoustic cues to vowel categories, though not uniformly informative across the vowel space (at least not in the Peterson and Barney vowel corpus).

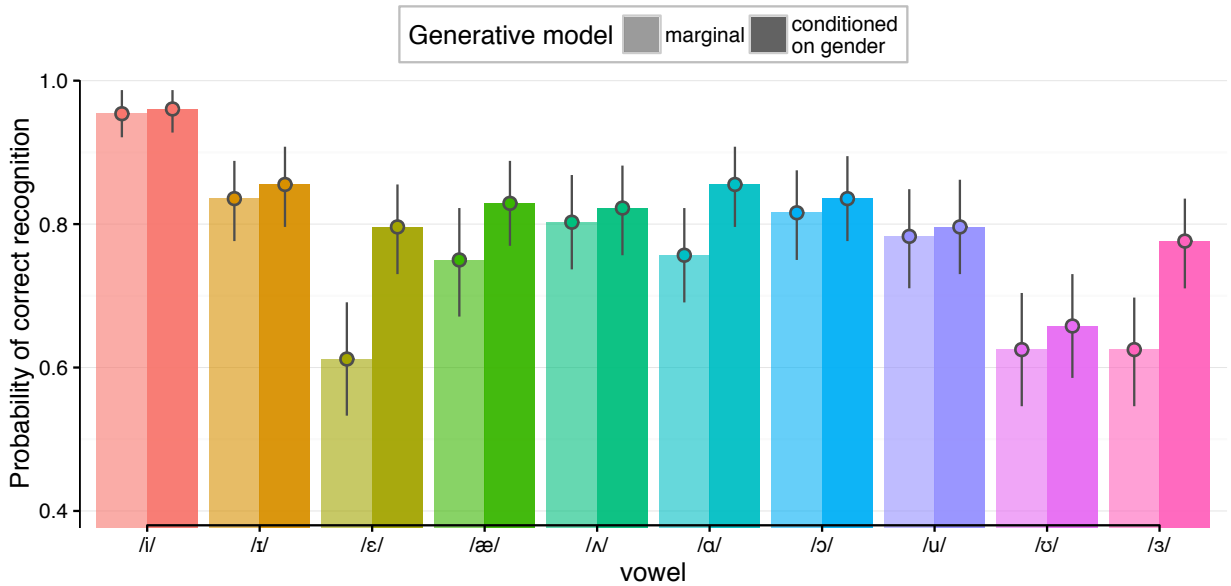

Figure 7: Probability of correctly recognizing vowel tokens under an ideal observer model with perfect knowledge of the gender-specific F_1xF_2 distributional statistics for each vowel in the Peterson and Barney (1952) vowel corpus (dark bars), versus an ideal observer model with only knowledge of the marginal F_1xF_2 distributions for each vowel (light bars).

The gender-specific and talker-specific analyses presented above lead to questions about the relative information gender or talker provide about the distributional statistics of vowels. Does conditioning the cue-to-category mapping on both gender *and* talker improve category recognition beyond conditioning on gender alone? Or perhaps it is the case that conditioning on gender provides a particularly *efficient* way of improving category recognition (e.g., learning and storing talker-specific distributional statistics will subsume gender-specific statistics but with more degrees of freedom; i.e., more to learn and store). A comparison between Figures 5 and 7 might suggest that the two socio-indexical variables carry (partially) complementing information about vowel statistics. However, differences between the vowel corpora that we have used here (Heald and Nusbaum, 2015; Peterson & Barney, 1952) prevent us from making this conclusion. The two corpora differ along a number of critical dimensions (number of vowel categories, number of talkers of each gender, number of vowel tokens per talker), which makes it infelicitous to compare the relative size of the category recognition differences between Figure 5 and Figure 7.[7] However, we show in Section 4 how one can address this type of question in future research.

So far we have shown that knowledge of the covariance between linguistic categories and social categories can markedly improve the accuracy of speech perception. Next we discuss how ideal observer models can be extended to capture this intuition.

2.3. The ideal adapter framework

The ideal adapter framework for speech perception builds on the logic of ideal observer models, but extends these models to account for the fact that the statistics of the world are subjectively non-stationary. The foundation of the ideal adapter framework (Kleinschmidt & Jaeger, 2015b) is that listeners cope with

[7] We used the Heald and Nusbaum (2015) for the talker-specific analyses because it comprises a large number of vowel tokens from multiple talkers and, hence, affords reliable estimates of talker-specific distributional statistics. However, since the Heald and Nusbaum (2015) corpus contains productions from only 5 females and 3 males, this corpus was not suitable for reliably assessing gender-specific distributional statistics. By contrast, the Peterson and Barney (1952) corpus comprises productions from a large number of males and females, though only two tokens of each vowel from each talker. Hence, the Peterson and Barney (1952) corpus is well suited for estimating gender-specific distributional statistics across talkers, but not for estimating talker-specific distributions.

uncertainty in speech perception by adapting their beliefs about the distribution of acoustic-phonetic cue values associated with speech categories as they experience new input. When listeners' beliefs change—e.g., due to exposure to new talkers or to an unfamiliar dialect or sociolect—so do their inferences (posterior probabilities), which are based on those beliefs. **Figure 8** illustrates this way of reasoning.

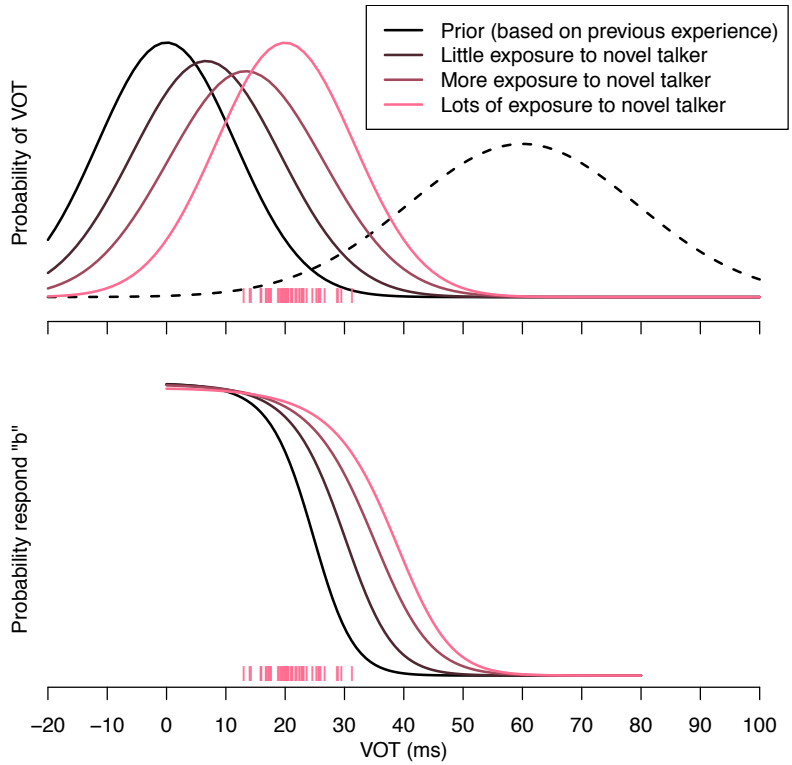

Figure 8: Schematic illustrations of phonetic adaptation. The top panel shows hypothetical VOT distributions for the categories /b/ (solid lines) and /p/ (dashed line). The bottom panel shows the ideal classification function for /b/ based on those distributions. The black lines in the top panel of Figure 4 show hypothetical prior beliefs: note that the prior belief distribution for /b/ is centered at 0ms, and the point of maximal ambiguity between /b/ and /p/ is ~25ms. The red rug of data points between 15ms and 35ms indicates recently experienced tokens of /b/ that were acoustically ambiguous between /b/ and /p/. As the listener accumulated experience with these tokens (e.g., hearing the word *breakfast* pronounced like "[b/p?]reakfast"), the listener updates her beliefs about the distribution of VOTs associated with /b/, as indicated by the differently colored lines in the top panel. As a result of this belief updating, the ideal classification function derived from those beliefs shifts such that the otherwise acoustically ambiguous tokens are unambiguously associated with /b/, as shown by the colored lines in the bottom panel. For a model that implements such changes in beliefs about the likelihood distribution and a test of this model, see Kleinschmidt and Jaeger (2015b).

Two aspects of Kleinschmidt and Jaeger's (2015b) model are crucial. First, they proposed that listeners *learn* and *continuously update* their beliefs about the variance-covariance between linguistic variability and the factors that systematically condition this variability, such as talker identity and social group. Second, Kleinschmidt and Jaeger (2015b) proposed that the speech comprehension system *stores* learned (adapted) beliefs about relevant distributions so that the information can later be retrieved and used in similar contexts (see also Foulkes & Docherty, 2006; Hay, Warren, & Drager, 2006; Pierrehumbert, 2001; Johnson, 1997). For instance, it can store acoustic characteristics of phonemes produced by a familiar talker (e.g., a friend, family member, celebrity) or a member of a talker group (e.g., gender, age, dialect, accent). Storing learned distributional information can be advantageous because it makes it unnecessary to begin the adaptation process anew each time a familiar talker is encountered. Rather, by storing learned distributional information, listeners can reuse this information to recognize speech from familiar talkers and to generalize to new similar talkers. In other words, given a

social cue that correlates with a pattern of linguistic variation, listeners can leverage socio-indexical information to infer the intended cue-to-category mapping. For instance, one can view the talker's gender as an important conditioning factor for the distribution of spectral energy associated with certain sound categories (see Figure 6; Strand & Johnson, 1996; Hillenbrand et al., 1995; Peterson & Barney, 1952), and can leverage gender-specific distributional information to guide the cue-to-category mapping by constraining the estimates of the likelihood and prior. This reasoning is expressed formally as follows:

(3) $p(\text{C}=\text{c}_i \mid \text{acoustic cues, socio-indexical cues}) \propto p(\text{acoustic cues} \mid \text{C}=\text{c}_i, \text{socio-indexical cues}) * p(\text{C}=\text{c}_i, \text{socio-indexical cues})$

In summary, the ideal adapter framework makes the following predictions.

1) *Inference of linguistic categories:* Listeners infer phonological categories by combining their prior expectations for a category with their beliefs about an underlying model that generated the observed acoustic data. These inferences are conditioned on contextual factors, including social-indexical features of talkers and talker groups, that predict distributions of acoustic cue values for a given phoneme category.
2) *Inference of generative model:*
   a. Listeners learn to adapt their beliefs as they obtain more data.
   b. Listeners store the learned distributions of the acoustic cue values so as to reuse them when they encounter a similar context in the future.

To what extent do existing empirical data support the predictions that listeners learn, store and utilize information about how statistical properties of speech *covary* with social-indexical features? We turn to this issue in the next section.

3. Evidence of speech perception conditioned on socio-indexical linguistic knowledge

A growing body of evidence demonstrates that the speech perception system is sensitive to the statistical properties of speech. A now well-attested finding is that short-term exposure to statistical distributions that differ from the long-term statistics of the language (e.g., repeated exposure to atypically realizations of phonetic contrasts) causes listeners to shift their category boundaries (classification functions). For example, in a classic study by Norris et al. (2003), participants initially heard a sound that was acoustically ambiguous between [s] and [f] in either /s/-biased lexical contexts (e.g., "platypus[$?_{sf}$]" for *platypus*, an /s/-final word with no /f/-final counterpart) or /f/-biased lexical contexts (e.g., "gira[$?_{sf}$]" for *giraffe*, an /f/-final word with no /s/-final counterpart). As a result of this exposure, participants shifted their /s/-/f/ category boundary so the otherwise acoustically ambiguous stimulus was interpreted as /s/ by participants in the /s/-biased condition and as /f/ by participants in the /f/-biased condition. Similar phonetic recalibration after exposure to shifted pronunciations has been documented for a number of different phonetic contrasts and using different types of paradigms (e.g., Bertelson, Vroomen, & de Gelder, 2003; Eimas & Corbit, 1973; Kraljic & Samuel, 2005, 2006, 2007; McQueen, Norris, & Cutler, 2006; Maye, Werker, & Gerken, 2002; Miller, Connine, Schermer, & Kluender, 1983; Vroomen, van Linden, de Gelder, & Bertelson, 2007). More recent work suggests that this dynamic recalibration of category boundaries is indeed, at least in part, due to changes in beliefs about the underlying cue distributions, as predicted by the ideal adapter (e.g., the likelihood distribution, Clayards et al., 2008; Bejjanki et al., 2010, Kleinschmidt & Jaeger, 2015b; for further discussion, see Kleinschmidt & Jaeger, 2015a). This lends credibility to the ideas outlined in Section 2.

Our primary interest here, however, is to what extent speech perception is conditioned on socio-indexical linguistic knowledge. To this end, we discuss (i) whether listeners *learn* talker-specific and group-specific distributional statistics that characterize phonetic variability, and (ii) whether listeners *store* talker- and group-specific information for subsequent use.

3.1. Talker-specific and group-specific learning

Two conditions must be met for learning to be characterized as talker-specific. First, listeners must draw on experience with talker-specific distributional statistics to generate linguistic expectations that generalize to future utterances from that talker (i.e., learning the distributional statistics that characterize familiar category exemplars but that are independent of, or more precisely abstractions over, those exemplars). Second, listeners must fail to generalize these expectations to utterances from different talkers. Similarly, for learning to be characterized as *group-specific*, listeners must generate linguistic expectations that generalize to new talkers within a particular social group (*cross-talker within-group generalization*), but fail to generalize to talkers outside that group (*cross-talker cross-group generalization*).

A number of studies over the last two decades provide evidence that listeners learn talker-specific patterns of variability. In one early study, Nygaard and Pisoni (1998) trained participants to identify 10 unfamiliar talkers by voice and then tested the influence of talker identification on the recognition of novel words in noise. Participants in the experimental condition heard the same talkers during training and test, while control participants heard novel talkers at test. Overall, identification accuracy for the noise-masked words at test was significantly higher when participants were listening to the trained talkers than to the novel talkers (see also Nygaard, Sommers, & Pisoni, 1994). This finding cannot be attributed to memory for specific exemplars because the test words did not occur during training. Thus, participants learned abstract properties of the trained talkers' speech and drew on this knowledge to guide the mapping of speech input to linguistic categories. Nygaard and Pisoni (1998) did not assess whether this talker-specific learning was specifically distributional. However, the finding that participants learned abstract talker-specific information is consistent with the general predictions of the ideal adapter framework.

More recent studies further show that phonetic recalibration can be talker-specific (Eisner & McQueen, 2005; Kraljic & Samuel, 2005, 2007). In these studies, exposure to a particular talker led listeners to apply these recalibrated boundaries when subsequently listening to the trained talker but not to different talkers.

A second source of evidence that learning can be talker-specific is that when listening to multiple talkers, listeners can simultaneously learn that the same pronunciation variant maps to different categories depending on the talker (e.g., that [e^j] is one talker's realization of /e/ but another talker's raised realization of /ae/; Trude & Brown-Schmidt 2012), or that different pronunciation variants from different talkers map to the same category (e.g., that one talker has an atypically short VOT for /p/, while another talker has an atypically long VOT for /p/; Munson, 2011; Theodore & Miller, 2010).

This is not to say that learning *has to be* talker-specific: there is now suggestive evidence that generalization to another talker is blocked if talkers are perceived to be dissimilar with regard to the relevant phonetic contrast, but occurs when talkers are perceived to be similar (e.g., Reinisch & Holt, 2014; Kraljic & Samuel, 2007). Both talker-specificity and cross-talker generalization based on inter-talker similarity suggest that perceptual learning is sensitive to socio-indexical structure, as hypothesized by the ideal adapter, as well as by exemplar-based (e.g., Foulkes & Hay, 2015; Johnson, 1997), episodic (see e.g., Goldinger, 2007) and certain normalization accounts (e.g., McMurray & Jongman, 2011; Huang & Holt, 2009).

Adaptation to phonetic variability is not limited to talker-specific pronunciation patterns. Listeners can also abstract over individual talkers to learn patterns of pronunciation variation that are associated with particular groups of talkers (e.g., dialect or accent variation). Bradlow and Bent (2008) investigated the influence of exposure conditions on the specificity of adaptation to an unfamiliar foreign accent. Participants who only heard speech from a single foreign-accented talker adapted talker-specifically (i.e., no generalization to a new talker with the same accent). By contrast, participants who heard speech from multiple talkers with the same foreign accent were able to abstract over talker-specific pronunciation patterns to learn talker-independent accent-specific variation (i.e., generalization at test to a new talker

with the same accent, but not to a new talker with a different accent; see also Best et al., 2015; Sidaras et al., 2009; Xie, Theodore, & Myers, submitted).[8]

The scope of group-specific adaptation does not seem to be bound to a particular level of group categorization. In a follow-up study, Baese-Berk et al. (2013) demonstrated that exposure to multiple foreign accents provided an advantage in adaptation to other novel foreign accents. In other words, the relevant "group" in Baese-Berk et al.'s (2013) study was not talkers with a particular accent (e.g., Mandarin-accented talkers; cf. Bradlow & Bent, 2008), but rather accented talkers more generally.

3.2. Talker-specific and group-specific storage

A separate line of studies has provided evidence that listeners store information about characteristics of individual talkers and apply this knowledge in real-time speech processing. Research within the exemplar-based tradition, for example, has shown that listeners encode fine-grained phonetic detail of speech episodes from the talkers they encounter, and further that memory of talker-specific episodic detail facilitates recognition when familiar episodes are encountered again (e.g., when hearing the same word repeated by a given talker; Bradlow et al., 1999; Palmeri et al., 1993). Goldinger (1996) investigated the longevity of such recognition effects by varying the interval between the first and repeated instances of the talker-specific episodes. The recognition benefit occurred even after 1-week (the longest interval tested), which indicates that the episodic detail was stored in long term memory (see also Goldinger & Azuma, 2004). Sensitivity to talker-specific pronunciation information is not limited to previously encountered episodes (e.g., word forms). With repeated exposure to a particular talker, listeners can build abstract talker-specific representations of phonetic categories: that is, listeners can *adapt* abstract category representations to reflect the distribution of acoustic cues associated with a talker's category realizations (Reinisch & Holt, 2014; Kraljic & Samuel 2005, 2007; Eisner & McQueen, 2005). By building talker-specific representations of phonetic categories, listeners are able to *generalize* learning to new words that contain talker-specific category variants (Sjerps & McQueen, 2010; McQueen et al., 2006). Eisner and McQueen (2006) found that the phonetic category adjustments resulting from short-term exposure to a particular talker persisted for at least 12 hours (the longest interval tested), which suggests that abstract talker-specific category representations are stored in memory (see also Earle & Myers, 2015; Nygaard & Pisoni, 1998).

In addition to stored talker-specific knowledge, listeners seem to use stored group-specific social-indexical knowledge—such as information about a talker's age, gender or regional background—to guide speech perception (Drager, 2011; Walker & Hay, 2011; Hay, Warren & Drager, 2006; Strand & Johnson, 1996). For example, Niedzielski (1999) found that listeners' expectations about a talker's regional background affected perception of target vowel tokens. The same raised-/aw/ token (e.g., *about* pronounced more like "a boat") was heard as raised by Detroit participants who were told the talker was from Canada, but heard as unraised by Detroit participants who were told the talker was from Michigan. Thus, top-down instructions about the talker's social background seem to affect how listeners processed the bottom-up input. Niedzielski reasoned that this effect was driven by stereotypes about Canadian English: Detroiters expect to hear a raised /aw/ in the speech of a Canadian talker, which makes it more likely for them to actually *hear* it in the stimuli. On the other hand, they do not expect to hear that variant in their own variety, and hence they do not. In a related study, concerning vowel variation in New Zealand English, Hay et al. (2006) used pictures to manipulate the perceived age and social class of a set of New Zealand talkers and found that participants expectations about the talkers' social attributes affected vowel perception in a manner consistent with the age- and class-based variation. In order for

[8] Intriguingly, in the learning of non-native phonetic contrasts, cross-talker generalization based on single-talker exposure seems to occur only after sleep, perhaps due to integration of the learned talker-specific statistics with prior knowledge (Earle & Myers, 2015). This might suggest that cross-talker generalization only when listeners already had prior implicit knowledge about similar talkers. In those cases, exposure to the single talker might allow listeners to reduce uncertainty about which type of talker model to apply in the current situation (cf. Kleinschmidt & Jaeger, 2015b: p. 180-182).

these effects to occur in the absence of recent exposure, listeners must have stored knowledge about the variance-covariance between linguistic information and social groups.

4. When should listeners adapt?

The premise of the ideal adapter framework is that having the right expectations facilitates language processing while having the wrong ones impairs it. For example, to the extent that individual talkers differ in their linguistic behavior from the population-level distributional statistics, then deriving expectations about the appropriate cue-to-category mapping for a given utterance based on population-level statistics provides a poor fit to data from individual talkers (see, e.g., Figure 2 above). It is thus beneficial for listeners to adapt to and store information about systematic variability in order to derive more precise predictions about how variable input relates to linguistic categories. However, there is potentially an infinite number of covariates (conditioning factors) that *could* explain variability in the speech signal. Adaptation to and storage of covariances are presumably associated with a non-zero cognitive cost, raising questions about the utility of adaptation. Even if the human brain was able to learn and store distributional information for all covariates of a given linguistic phenomenon, the computational tractability of estimating conditional probabilities decreases as the number of covariates on which those probabilities are estimated increases. How, then, does the brain determine which covariance statistics to rely on in order to achieve robust comprehension *efficiently*? For example, is there a principled reason *why* listeners seem to be exquisitely sensitive to talker- and group-specific variability? In the remainder of this article, we illustrate how an ideal observer approach can help us to begin to conceptualize this question more clearly. Specifically, we illustrate how the ideal observer approach allows us to quantify the utility of learning and storing specific covariance statistics and to assess the relative informativity of multiple covariance statistics.

4.1. When are covariance statistics useful?

Several factors determine the utility of learning and storing information about the covariance between, on the one hand, a specific non-acoustic cue (such as talker identity), and, on the other hand, category-specific distributions of acoustic cues. First, there are general considerations about the utility of correct classification of a linguistic category. This utility presumably depends on how relevant correct classification is for the current purpose. For example, if the current task is to correctly classify a category the utility of doing so is high. If, however, the comprehender's current goal is to successfully infer the message intended by the speaker, correct categorization of a phonetic category is relevant only to the extent that correct recognition of this category facilitates successful recognition of the intended *message* (for related discussion, see Kuperberg & Jaeger, 2016). Beyond these general considerations, the utility of learning and storing the covariance between the posterior distribution of a linguistic category in Equation (1) and a specific contextual cue (e.g., talker identity) also depends on how much doing so facilitates correct categorization.[9] This in turn depends on the amount of information the contextual cue carries about the linguistic category distribution, as well as the certainty with which the contextual cue can be correctly recognized on average (e.g., certainty about talker identity).

One advantage of the ideal observer approach is that it allows us to quantify this information. In Section 2.2.2, we assessed the informativity of talkers with respect to vowel categories. Here, we expand this approach to directly compare the (relative) informativity of multiple contextual factors with respect to the same cue and category distributions. We again use the vowel corpus collected by Heald and Nusbaum (2015). Recall that in this corpus, each talker produced multiple tokens of each vowel category three times a day (at 9am, 3pm and 9pm) on multiple different days. In addition to cross-talker variability in

[9] The same logic can be applied to the phonetic cues themselves. The reason that listeners condition their perception of, for example, voicing contrasts on a specific feature is presumably that this feature is informative about category identity, thus facilitating correct category classification (cf. Toscano & McMurray, 2010; Vallabha, McClelland, Pons, Werker, & Amano, 2007).

vowel formants (see Figure 3), Heald and Nusbaum found that some formants (*f*0 and F_1, though not F_2 and F_3) showed significant, though small, differences throughout the day (likely due to the effects of articulatory fatigue on vowel productions). Thus, an ideal observer with perfect knowledge of both talker-dependent *and* time-of-day-dependent variability could derive more accurate inferences about the cue-to-category mapping than an ideal observer that lacked knowledge of one or both of these sources of variability. The question arises, however, of whether time-of-day is a sufficiently informative contextual cue to warrant the cost of learning and storing time-of-day-specific distributional statistics.

We can assess the informativity of contextual cues at multiple levels. As in Section 2.2.2, we first assess informativity at the level of cue distributions. For this, we calculated the KL divergence between the marginal F_1xF_2 distribution for each vowel category in Heald and Nusbaum's corpus—*p*(category | F_1xF_2)—and the corresponding distribution conditional on various contextual factors: i.e., conditional on time of day—*p*(category | F_1xF_2, time of day); conditional on talker identity—*p*(category | F_1xF_2, talker); or conditional on both time of day *and* talker—*p*(category | F_1xF_2, time of day, talker). Figure 9 shows the results of this comparison (see Appendix A for further details. Note that the values for the KL divergence between the marginal and talker-dependent distributions—middle bars—are the same values shown in Figure 4 and are repeated here for comparison). Figure 9 shows that the average information loss of ignoring time of day (leftmost bars) is effectively 0 bits: that is, the marginal distributional statistics provide a good fit to the time-of-day-dependent distributional statistics. Against this baseline, the information loss of ignoring talker identity is high, which suggests that there is considerable information to be gained from learning and storing talker-specific distributional statistics. However, jointly conditioning on time-of-day and talker identity adds minimal if any information about cue distributions beyond conditioning on talker identity alone (rightmost bars in Figure 9). Thus, this cue-level analysis suggests that time-of-day is *not* sufficiently informative to warrant learning and storing time-of-day-dependent distributional statistics.

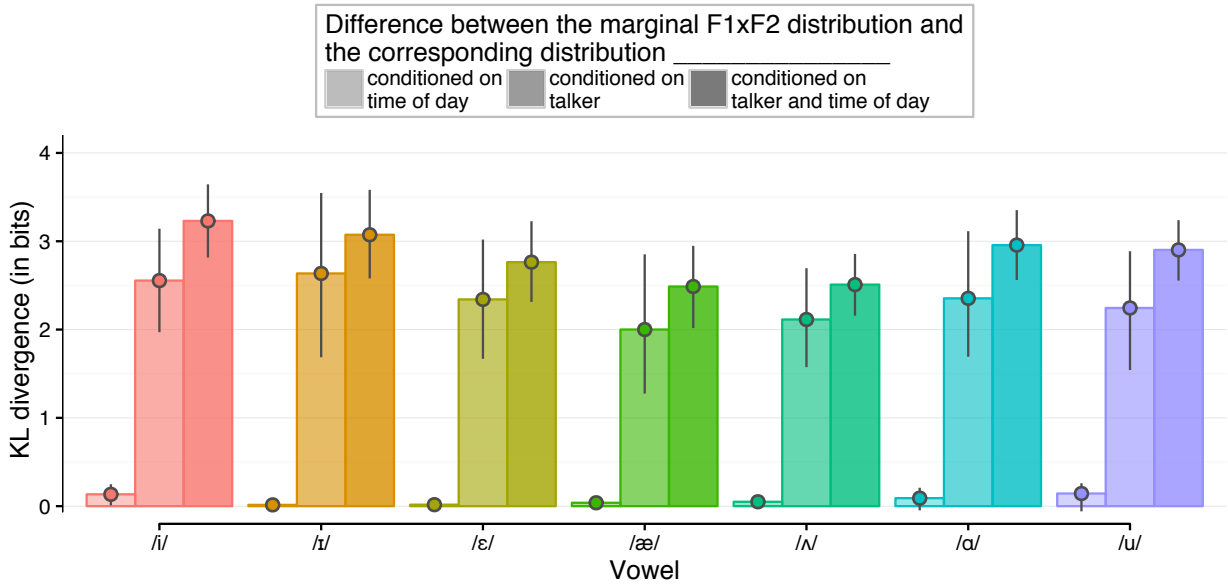

Figure 9. Information loss when ignoring various factors in inferring vowel identity, plotted separately for each vowel. Points indicate the KL divergence between the marginal F_1xF_2 distribution in Heald and Nusbaum's (2015) corpus, on the one hand, and the corresponding time-of-day-dependent distribution (orange dots), talker-dependent distribution (green dots), and talker-by-time-of-day-dependent distribution (blue dots) on the other hand. There is minimal, if any, information loss when ignoring time-of-day-dependent variability. By contrast, there is substantial information loss when ignoring talker identity (i.e., inferring vowel identity based on population-level rather than talker-dependent statistics).

We can also assess the informativity of contextual cues at the level of category recognition. That is, as already shown in Section 2.2.2, we can directly assess the effect that learning and storing conditional

distributional statistics is predicted to have on the rate of correct categorizations. To compare the informativity of various contextual factors, we calculated the change in ideal category recognition accuracy depending on which of the following distributional statistics were available to the ideal observer: the marginal F_1xF_2 distributional statistics for each vowel; or the corresponding distributional statistics conditional on either time of day or talker identity, or conditional on both time of day and talker identity. Figure 10 shows the results of this comparison (note that recognition accuracy under the marginal and talker-specific models is the same as reported in Figure 5 and is repeated here for comparison). Ideal vowel recognition accuracy was quite low (between 45% and 80% correct) when the cue-to-category mapping was based on the marginal distribution—p(category | F_1xF_2)—with the notable exceptions of the point vowels /u/ and /i/, for which recognition accuracy was effectively at ceiling. Conditioning on time of day—p(category | F_1xF_2, time of day)—yielded minimal, if any, improvement in recognition accuracy, relative to the marginal model (consistent with the finding that the KL divergence between the marginal and time-of-day-dependent distributions is effectively zero). Conditioning the cue-to-category mapping on talker identity—p(category | F_1xF_2, talker)—yielded a substantial increase in recognition accuracy relative to the marginal model (upwards of 40% for some vowels). By contrast, jointly conditioning on time-of-day and talker identity provided no consistent improvement in recognition accuracy beyond conditioning on talker identity alone (with the apparent exception of the vowel /æ/).

Assessing informativity at the level of category recognition provides further evidence that learning and storing talker-specific distributional statistics is useful, whereas time-of-day-specific distributional statistics lack utility. The latter is not surprising: although Heald and Nusbaum (2015) found statistically reliable formant differences at different time points throughout the day, these differences were small and, as they argue, likely below the *just noticeable difference* threshold and, hence, unavailable to listeners as a source of information.

We have shown that the utility of learning and storing conditional distributional statistics can be assessed computationally at multiple levels. A question for future research concerns the level(s) at which listeners assess information utility. The ultimate goal of speech processing is to understand the linguistic *message*; hence, information utility is perhaps best assessed in terms of the influence on comprehension (e.g., word recognition and pragmatic understanding). Another venue for future research, as we mentioned above, is to employ the present approach to compare whether the *relative* informativity of socio-indexical cues (for example, gender and talker-identity, or time-of-day) indeed predicts the extent to which human listeners learn and store covariance between these cues and phonological categories.

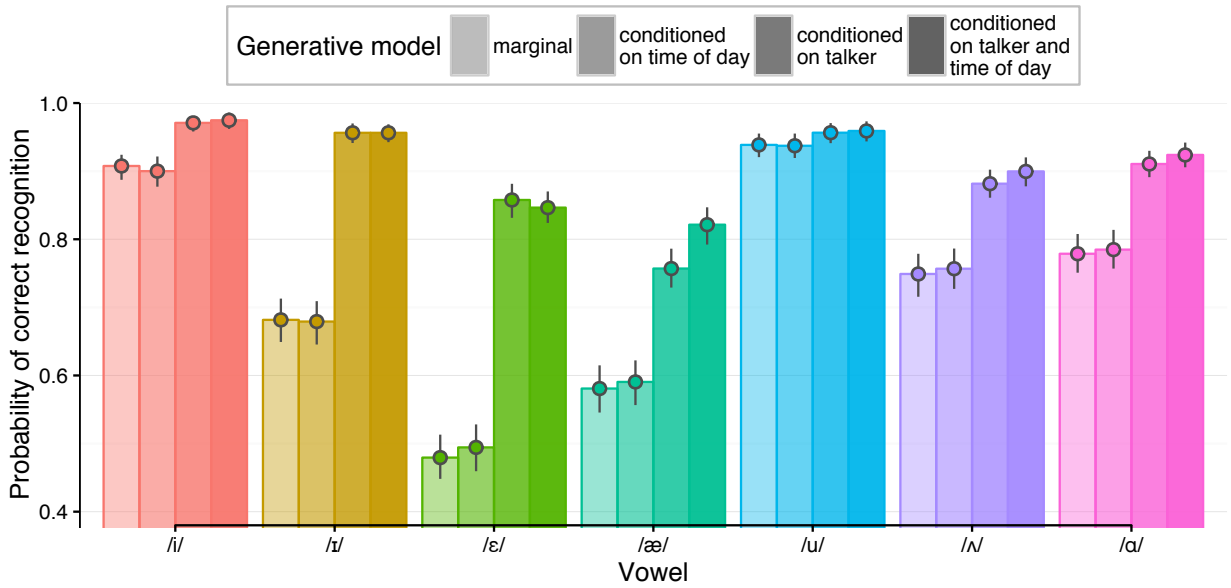

Figure 10. Average probability of correctly recognizing a vowel token as a function of different ideal observer models: marginal = p(category | F_1xF_2); conditioned on time of day = p(category | F_1xF_2, time); conditioned on talker = p(category | F_1xF_2, talker); conditioned on talker and time of day = p(category | F_1xF_2, talker, time of day).

5. Conclusion

One of the fundamental goals of research on speech perception is to explain how listeners reliably extract stable and accurate linguistic percepts from the speech signal despite the fact that the acoustic properties of speech are highly variable (lack of invariance). A growing body of research has demonstrated that the speech perception system is sensitive to statistical contingencies in how linguistic variability covaries with socio-indexical factors. Listeners seem to draw on this implicit statistical knowledge to guide the mapping of acoustic cue values to linguistic categories. Here our goal has been to demonstrate why Bayesian ideal observer-type models provide a framework within which sensitivity to such statistical contingencies can be productively understood. We have illustrated how the classic lack of invariance problem can be understood within the ideal observer approach and how we can use this approach to understand *why* listeners are so exquisitely sensitive to socio-indexical cues—because certain socio-indexical cues are highly informative about the mapping from acoustic cues to linguistic categories and, hence, conditioning inferences about linguistic categories on these socio-indexical cues *substantially* improves the probability of correct categorization. In illustrating how these benefits can be quantified, we hope to also have shown how an ideal observer account can inform what type of contextual cues linguistic expectations are predicted to depend on.

As discussed above, many of the predictions of the ideal adapter framework are shared with episodic/exemplar-based models of speech perception (Pierrehumbert, 2006, 2002; Goldinger, 1997; Johnson, 1997) and with certain normalization accounts (McMurray & Jongman, 2011; Cole et al., 2010; Huang & Holt, 2009; Holt, 2005). It should be noted that the Bayesian view outlined here is not fundamentally at odds with these other approaches. A connection to exemplar approaches is straightforward: the conditional distributional statistics that form the basis of the ideal adapter model can be seen as derived from the exemplars that the observer has experienced. There is now considerable evidence that our knowledge is at least in part a (distributional) abstraction over exemplars (Pierrehumbert, 2001, 2002) One question for future research is whether listeners indeed learn and store conditional (talker- or group-specific) distributional statistics, as predicted by the ideal adapter framework, or whether these statistics are only derived online at decision time based on the distribution of

activated exemplars, as claimed by prominent exemplar-based models (for further discussion, see Kleinschmidt & Jaeger, 2015b).

The line of research discussed here further highlights the importance of understanding the statistics of the input when reasoning about speech perception (Kleinschmidt & Jaeger, 2015a; Feldman et al., 2009; Clayards et al., 2008; Norris & McQueen, 2008; McClelland & Elman, 1986). Similar reasoning extends to language processing beyond perception (Smith & Levy, 2013; Levy, 2008; Altmann & Kamide, 1999; MacDonald, Pearlmutter & Seidenberg, 1994; Spivey-Knowlton, Trueswell & Tanenhaus, 1993; see Kuperberg & Jaeger, 2016, for recent overview): rather than to solely rely on marginal statistics, as is still assumed in most of the literature on processing above the level of the spoken word, a robust comprehension system might need to learn and store information about the covariance between variable linguistic behavior (e.g., lexical, syntactic or pragmatic variability) and the contextual factors that condition the variability (see also Yildirim, Degen, Tanenhaus, & Jaeger, 2016; Fine, Jaeger, Farmer, & Qian, 2013; Kamide, 2012; Kurumada, Brown, & Tanenhaus, 2012; Creel, Aslin, & Tanenhaus, 2008).

Appendix A. Measuring the difference between (multivariate) distributions

The *Kullback–Leibler (KL) divergence* is an information-theoretic measure of the difference between two distributions. Specifically, KL divergence, denoted $D_{KL}(P||Q)$, measures the information (in bits) that is lost when using one probability distribution (Q) to estimate another (P):

(4) $$D_{KL}(P \,||\, Q) = \int \left[\log(p(x)) - \log(q(x))\right] p(x)\, dx$$

To determine the informativity of talkers with regard to acoustic cue distributions, we measured the KL divergence between the population-level distribution of vowel categories given an observed F_1 and F_2 value, *p*(category | F_1xF_2), and the same category distribution additionally conditioned on talker identity, *p*(category | F_1xF_2, talker)—under the non-critical assumption that vowel categories form multivariate normal distributions in formant space. The KL divergence between two multivariate Gaussians has an analytic solution (Ramey, 2013), which for the current case of vowel distributions is:

(5) $$D_{KL}(P = p(\text{vowel} \mid \text{acoustic cues}, \text{talker}) \,||\, Q = p(\text{vowel} \mid \text{acoustic cues}) = \frac{1}{2}\left[\log\frac{|\Sigma_2|}{|\Sigma_1|} - d + \text{tr}(\Sigma_2^{-1}\Sigma_1) + (\mu_2 - \mu_1)^T \Sigma_2^{-1}(\mu_2 - \mu_1)\right]$$

where *P* and *Q* are multivariate normal distributions (over e.g., F_1 x F_2); μ_1 and Σ_1 are the mean and variance-covariance matrix of the vowel under P; and μ_2 and Σ_2 are the mean and variance-covariance matrix of the vowel under Q; *d* is the dimensionality of the variance-covariance matrix (which is the same for P and Q; e.g., 2 for the case of F_1 x F_2); *tr* is the trace of a matrix (sum of diagonal elements); and *T* is the transpose.

If talker-dependent distributions on average differ considerably from the population-level distribution, then the information loss (and thus the KL divergence) between the talker-dependent and population-level distributions is high.[10] By contrast, if the KL divergence between the population level and talker-dependent distributions is low, then the population-level statistics provide a reasonably good fit to any talker-specific data. Hence, in the latter case, there is little information to be gained by learning and storing talker-specific distributional information.

[10] Here, we continue to use the term talker-*specific* when talking about listeners' expectations or probabilistic beliefs and the idea that they are conditioned on talker-identity. We use the term talker-*dependent* when referring to the distributions observed in the world or speech signal.

Appendix B. Modeling category recognition accuracy

We briefly describe the procedure we used for modeling statistically optimal category recognition accuracy. We first calculated the posterior probability of each vowel category given the F_1 and F_2 of each vowel token in the vowel corpora. That is, for each pair of observed F_1 and F_2 values, we calculated the posterior probability of /i/, /ɪ/, etc. For the Figures presented in the main text, we then selected the vowel category with the highest posterior probability as the category to which the ideal observer model mapped the input (applying the *criterion rule*, Friedman & Massaro, 1998). Similar results were obtained when applying *Luce's choice rule* (Luce, 1959), which chooses vowel responses proportional to their posterior probability. We repeated this procedure for multiple ideal observers: i.e., an ideal observer with access to only the marginal (population-level) F_1xF_2 distributional statistics for each vowel, and an ideal observer with access to the corresponding distributional statistics conditional on talker identity, talker gender, and/or time of day.

For these analyses, we made the following assumptions. First, we assumed uniform category priors (i.e., that each vowel was equally likely to occur, based on previous experience). Thus, the posterior probability of category $C = c_i$, given the observed F_1 and F_2 values, was proportional to the corresponding likelihood term—$p(F_1xF_2 \mid C = c_i)$. This assumption can easily be relaxed by weighting the likelihood term by the corresponding category prior. Second, we assumed that vowel category representations can be approximated as multivariate normal distributions in F_1xF_2 (Hz) space. This assumption can easily be adjusted to include additional cues to vowel identity (e.g., $F_1xF_2xF_3$) or vowel cues in a normalized or transformed vowel space (Bark, Lobanov). Third, we assumed that vowel categories form *multivariate Normal distribution* in F_1xF_2 space. This assumption could be relaxed by using, for example, kernel density estimates of the distributions. The assumption of normality likely underestimates the overall performance of the ideal observer (which now would be free to capture non-normally distributed cue distributions). Our first three assumptions do not introduce a trivial bias into our analyses of the impact of talker identity on vowel recognition.

Fourth, for the talker-specific ideal observer model, we assumed perfect certainty about talker identity. Fifth and finally, we derived the talker-specific statistics that the talker-specific ideal observer uses separately from each talker's data, without assuming any prior beliefs about the distribution of category-specific cue distributions across talkers. These estimates are thus likely overestimating inter-talker differences, compared to the true underlying differences between talkers. Both the fourth and fifth assumption bias our estimation of the impact of talker-identity towards overestimation. For the present purpose, we consider this acceptable.

References

Adank, P., Smits, R., & van Hout, R. (2004). A comparison of vowel normalization procedures for language variation research. *Journal of the Acoustical Society of America*, 116(5), 3099-3107.

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

Allen, J. S., & Miller, J. L. (2004). Listener sensitivity to individual talker differences in voice-onset-time. *Journal of the Acoustical Society of America*, 115 (6), 3171-3183.

Altmann, G. T., & Kamide, Y. (1999). Incremental interpretation at verbs: restricting the domain of subsequent reference. *Cognition*, 73(3), 247-264. doi: 10.1016/S0010-0277(99)00059-1

Anderson, J. R. (1990). The adaptive character of thought. Hillsdale, NJ: Erlbaum.

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, 98(3), 409-429. doi: 10.1037/0033-295X.98.3.409

Baese-Berk, M. M., Bradlow, A. R., & Wright, B. A. (2013). Accent-independent adaptation to foreign accented speech. *Journal of the Acoustical Society of America*, 133 (3), 174-180. doi: 10.1121/1.478986

Bejjanki, V. R., Clayards, M., Knill, D. C., & Aslin, R. N. 2011. Cue integration in categorical tasks: Insights from audio-visual speech perception. *PLoS One*, 6 (5), e19812. doi: 10.1371/journal.pone.0019812

Best, C., Shaw, J., Mulak, K., Docherty, G., Evans, B., Foulkes, P., Hay, J., Al-Tamimi, J., Mair, K. & Wood, S. (2015), Perceiving and adapting to regional accent differences among vowel subsystems. Proceedings of the 18th *International Congress of Phonetic Sciences*, Glasgow, Scotland.

Bertelson, P., Vroomen, J., & de Gelder, B. (2003). Visual recalibration of auditory speech identification: A McGurk after effect. *Psychological Science*, 14, 592-597. doi: 10.1046/j.0956-7976.2003.psci_1470.x

Bicknell, K., Tanenhaus, M. K., & Jaeger, T. F. (under review). Listeners maintain and rationally update uncertainty about prior words in spoken comprehension.

Bradlow, A. R., & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition*, 106(2), 707-729. doi: 10.1016/j.cognition.2007.04.005

Bradlow, A. R., Nygaard, L. C., & Pisoni, D. B. (1999). Effects of talker, rate, and amplitude variation on recognition memory for spoken words. *Perception and Psychophysics*, 61(2), 206-219. doi: 10.3758/BF03206883

Brown-Schmidt, S., Yoon, S. O., & Ryskin, R. A. (2015). People as contexts in conversation. In B. Ross (Ed.), The psychology of learning and motivation. (59-99). Academic Press: Elsevier Inc.

Chater, N., & Oaksford, M. (Eds.). (1998). *Rationality in an uncertain world: Essays in the cognitive science of human understanding*. East Sussex, UK: Psychology Press. doi: 10.4324/9780203345955

Clayards, M., Tanenhaus, M. K., Aslin, R. N., & Jacobs, R. A. (2008). Perception of speech reflects optimal use of probabilistic speech cues. *Cognition*, 108(3), 804-809. doi: 10.1016/j.cognition.2008.04.004

Cole, J. S., Linebaugh, G., Munson, C., & McMurray, B. (2010). Unmasking the acoustic effects of vowel-to-vowel coarticulation: A statistical modeling approach. *Journal of Phonetics*, 38. 167–184, doi:10.1016/j.wocn.2009.08.004

Creel, S. C., Aslin, R. N., & Tanenhaus, M. K. (2008). Heeding the voice of experience: The role of talker variation in lexical access. *Cognition*, 106, 633-664. doi: 10.1016/j.cognition.2007.03.013

Drager, K. (2011). Speaker age and vowel perception. *Language and Speech*, 54, 99-121. doi: 10.1177/0023830910388017

Earle, F. S., & Myers, E. B. (2015). Overnight consolidation promotes generalization across talkers in the identification of nonnative speech sounds. *Journal of the Acoustical Society of America*, 137, EL91. doi: 10.1121/1.4903918

Eisner, F., & McQueen, J. M. (2005). The specificity of perceptual learning in speech processing. *Perception and Psychophysics*, 67(2), 224-238. doi: 10.3758/BF03206487

Feldman, N. H., Griffiths, T. L., & Morgan, J. L. (2009). The influence of categories on perception: explaining the perceptual magnet effect as optimal statistical inference. *Psychological Review*, 116(4), 752–82. doi: 10.1037/a0017196

Fine, A. B., Jaeger, T. F., Farmer, T. A., & Qian, T. (2013). Rapid expectation adaptation during syntactic comprehension. *PLoS One*, 8. doi: 10.1371/journal.pone.0077661

Fine, A. B., Qian, T., Jaeger, T. F., & Jacobs, R. A. (2010). Is there syntactic adaptation in language comprehension? In Proceedings of the *48th Annual Meeting of the Association for Computational Linguistics: Workshop on cognitive modeling and computational linguistics*. Uppsala, Sweden.

Fitch, W. T., and Giedd, J. (1999). Morphology and development of the human vocal tract: A study using magnetic resonance imaging. *Journal of the Acoustical Society of America*, 106(3). 1511-1522. doi: 10.1121/1.427148

Foulkes, P., & Hay, J. (2015). The emergence of sociophonetic structure. In B. MacWhinney and W. O'Grady (Eds.), *The handbook of language emergence* (292-313). Hoboken, NJ: John Wiley and Sons, Inc. doi: 10.1002/9781118346136.ch13

Foulkes P, & Docherty G. J. (2006). The social life of phonetics and phonology. *Journal of Phonetics*, 34(4), 409–38. doi:10.1016/j.wocn.2005.08.002

Friedman, D., & Massaro, D. W. (1998). Understanding variability in binary and continuous choice. *Psychonomic Bulletin & Review*, 5(3): 370.389. doi: 10.3758/BF03208814

Geisler, W. S. (2003). Ideal observer analysis. In L. Chalupa & J. Werner (Eds.), *The visual neurosciences* (825-837). Boston: MIT Press.

Goldinger, S. D. (2007). A complementary-systems approach to abstract and episodic speech perception. In Proceedings of the *16th International Congress of Phonetic Sciences*, ed. J Trouvain, W.J. Barry, pp. 49–54. Dudweiler, Germany: Pirrot

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review*, 105, 251-279. doi: 10.1037/0033-295X.105.2.251

Goldinger, S. D. (1996). Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22(5), 1166-1183. doi: 10.1037/0278- 7393.22.5.1166

Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. B. (2010). Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in Cognitive Sciences*, *14*(8), 357–64. doi:10.1016/j.tics.2010.05.004

Griffiths, T. L., Lieder, F., & Goodman, N. D. (2015). Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic. *Topics in Cognitive Science*, 7, 217-229. doi: 10.1111/tops.12142

Griffiths, T. L., Vul, E., & Sanborn, a. N. (2012). Bridging levels of analysis for probabilistic models of cognition. *Current Directions in Psychological Science*, *21*(4), 263–268. doi:10.1177/0963721412447619

Hay J., Warren, P., & Drager, K. (2006). Factors influencing speech perception in the context of a merger-in-progress. *Journal of Phonetics*, 34, 458-484. doi: 10.1016/j.wocn.2005.10.001

Hay, J., & Drager, K. (2007). Sociophonetics. *Annual Review of Anthropology*, 36, 89-103. doi: 10.1146/annurev.anthro.34.081804.120633

Hay, J., Pierrehumbert, J. B., Walker, A. J., & LaShell, P. (2015). Tracking word frequency effects through 130 years of sound change. *Cognition*, 139, 83-91. doi: 10.1016/j.cognition.2015.02.012

Heald, S. L. M., & Nusbaum, H. C. (2015). Variability in vowel production within and between days. *PLOS One*, 10(9). doi: 10.1371/journal.pone.0136791

Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *Journal of the Acoustical Society of America*, 97 (5), 3099-3111. doi: 10.1121/1.411872

Holt, L. L. (2005). Temporally nonadjacent nonlinguistic sounds affect speech categorization. *Psychological Science*, 16(4), 305-312. doi: 10.1111/j.0956-7976.2005.01532.x

Huang, J., & Holt, L. L. (2009). General perceptual contributions to lexical tone normalization. *Journal of the Acoustical Society of America*, 125(6), 3983-3994. doi: 10.1121/1.3125342

Howes, A., Lewis, R. L., & Vera, A. (2009). Rational adaptation under task and processing constraints: Implications for testing theories of cognition and action. *Psychological Review*, 116 (4), 717-751. doi: 10.1037/a0017187

Jacobs, R. A., & Kruschke, J. K. (2011). Bayesian learning theory applied to human cognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, 2(1), 8-21. doi: 10.1002/wcs.80

Kamide, Y. (2012). Learning individual talkers' structural preferences. *Cognition*, 124, 66-71. doi: 10.1016/j.cognition.2012.03.001

Kleinschmidt, D. F., Fine, A. B., & Jaeger, T. F. (2012). A belief-updating model of adaptation and cue combination in syntactic comprehension. Proceedings of the *34th Annual Conference of the Cognitive Science Society*, ed. by N. Miyake, D. Peebles, and R. P. Cooper, 599-604. Austin, TX.

Kleinschmidt, D. F., and Jaeger, T. F. (2015a). Re-examining selective adaptation: Fatiguing feature detectors, or distributional learning? *Psychonomic Bulletin & Review*. doi: 10.3758/s13423-015 -0943-z

Kleinschmidt, D. F., & Jaeger, T. F. (2015b). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review*, 122(2), 148-203. doi: 10.1037/a0038695

Kraljic, T., Brennan, S. E., & Samuel, A. G. (2008). Accommodating variation: Dialects, idiolects, and speech processing. *Cognition*, 107, 54-81. doi: 10.1016/j.cognition.2007.07.013

Kraljic, T., & Samuel, A. G. (2006). Generalization in perceptual learning for speech. *Psychonomic Bulletin and Review*, 13, 262-268. doi: 10.3758/BF03193841

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

Kraljic, T., & Samuel, A. G. (2005). Perceptual learning for speech: Is there a return to normal? *Cognitive Psychology*, 51. 141-178. doi: 10.1016/j.cogpsych.2005.05.001

Kurumada, C., Brown, M., & Tanenhaus, M. K. (2012). Pragmatic interpretation of contrastive prosody: It *looks like* speech adaptation. In Proceedings of the 34th annual meeting of the cognitive science society. Sapporo, Japan.

Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience*, *31*(1), 32-59.

Labov, W., Ash, S., & Boberg, C. (2006). *The atlas of North American English*. Berlin: Mouton de Gruyter.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, 106(3), 1126- 1177. doi: 10.1016/j.cognition.2007.05.006

Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech percepion revised. *Cognition*, 21(1), 1-36. doi: 10.1016/0010-0277(85)90021-6

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, 74(6), 431-461. doi: 10.1037/h0020279

Logan, J. S., Lively, S. E., & Pisoni, D. B. (1991). Training Japanese listeners to identify English /r/ and /l/: A first report. *Journal of the Acoustical Society of America*, 89(2), 874-876. doi: 10.1121/1.1894649

Luce, R. D. (1959). *Individual Choice Behavior: A Theoretical Analysis*. New York: Wiley.

MacDonald, M. C., Pearlmutter, N. J., & Seidenberg, M. S. (1994). The lexical nature of syntactic ambiguity resolution. Psychological Review, 101(4). 676-703. doi: 10.1037/0033-295X.101.4.676

Marr, D. (1982). Vision: A computational investigation into the human representation and processing of visual information. San Francisco: Freeman and Co.

Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, 82(3). B101-B111. doi: 10.1016/S0010-0277(01)00157-3

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, 18(1). 1–86. doi: 10.1016/0010-0285(86)90015-0

McQueen, J. M., Cutler, A., & Norris, D. (2006). Phonological abstraction in the mental lexicon. *Cognitive Science*, 30. 1113-1126. doi: 10.1207/s15516709cog0000 79

Miller, J. L., Connine, C. M., Schermer, T. M., & Kluender, K. R. (1983). A possible auditory basis for internal structure of phonetic categories. *Journal of the Acoustical Society of America*, 73 (6), 2124-2133. doi: 10.1121/1.389455

Munson, C. (2011). Perceptual learning in speech reveals pathways of processing (Unpublished doctoral dissertation. University of Iowa.

Newman, R. S., Clouse, S. A., & Burnham, J. L. (2001). The perceptual consequences of within-talker variability in fricative production. *Journal of the Acoustical Society of America*, 109(3). 1181-1196. doi: 10.1121/1.1348009

Niedzielski, N. (1999). The effect of social information on the perception of sociolinguistic variables. *Journal of Language and Social Psychology*, 18(1). 62-85. doi: 10.1177/0261927X99018001005

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology*, 47(2). 204-238. doi: 10.1016/S0010-0285(03)00006-9

Norris, D. (2006). The Bayesian reader: Explaining word recognition as an optimal Bayesian decision process. *Psychological Review*, 113 (2), 327-357.

Norris, D., & McQueen, J. M. (2008). Shortlist B: A Bayesian model of continuous speech recognition. *Psychological Review*, 115 (2), 357-395. doi: 10.1037/0033-295X.115.2.357

Norris, D., McQueen, J. M., & Cutler, A. (2015). Prediction, Bayesian inference and feedback in speech recognition. *Language, Cognition and Neuroscience*. doi: 10.1080/23273798.2015.1081703

Nygaard, L. C., Sommers, M. C., & Pisoni, D. B. (1994). Speech perception as a talker-contingent process. *Psychological Science*, 5(1), 42-46. doi: 10.1111/j.1467-9280.1994.tb00612.x

Nygaard, L. C., & Pisoni, D. B. (1998). Talker-specific learning in speech perception. *Perception and Psychophysics*, 60(3). 355-376. doi: 10.3758/BF03206860

Palmeri, T. J., Goldinger, S. D., & Pisoni, D. B. (1993). Episodic encoding of voice attributes and recognition memory for spoken words. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 19(2), 309-328. doi: 10.1037//0278-7393.19.2.309

Pardo, J. S., & Remez, R. E. (2006). The perception of speech. In M. Traxler & M. Gernsbacher (Eds.), *The handbook of psycholinguistics* (2nd ed., p. 201-248). Academic Press.

Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the vowels. *Journal of the Acoustical Society of America*, 24. 175-184. doi: 10.1121/1.1917300

Pierrehumbert, J. B. (2002). Word-specific phonetics. *Laboratory Phonology VII*, Mouton de Gruyter, Berlin, 101-139.

Pierrehumbert, J. B. (2006). The next toolkit. *Journal of Phonetics*, 34(4), 516–530. doi: 10.1016/j.wocn.2006.06.003

Pisoni, D. B. (1997). Some thoughts on 'normalization' in speech perception. In K. Johnson & J. W. Mullennix (Eds.), *Talker variability in speech processing* (9-33). San Diego: Academic Press.

Protopapas, A., & Lieberman, P. (1997). Fundamental frequency of phonation and perceived emotional stress. *Journal of the Acoustical Society of America*, 101(4), 2267-2277. doi: 10.1121/1.418247

Ramey, J. A. "KL divergence between two multivariate Gaussians." Stackexchange.com. Last modified June 3, 2013. http://www.stats.stackexchange.com/questions/60680/kl-divergence-between-two-multivariate-gaussians.

Reinisch, E., & Holt, L. L. (2014). Lexically-guided phonetic retuning of foreign-accented speech and its generalization. *Journal of Experimental Psychology: Human Perception and Performance*, 40 (2), 539-555. doi: 10.1037/a0034409

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2010). Rational approximations to rational models: Alternative algorithms for category learning. *Psychological Review*, *117*(4), 1144–1167. doi:10.1037/a0020511

Shi, L., Griffiths, T. L., Feldman, N. H., & Sanborn, A. N. (2010). Exemplar models as a mechanism for performing Bayesian inference. *Psychonomic Bulletin & Review*, *17*(4), 443–64. doi:10.3758/PBR.17.4.443

Sidaras, S., Alexander, J. E. D., & Nygaard, L. C. (2009). Perceptual learning of systematic variation in Spanish-accented speech. *Journal of the Acoustical Society of America*, 125 (5), 3306-3316. doi: 10.1121/1.3101452

Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological Review*, 63 (2), 129-138. doi: 10.1037/h0042769

Smith, N. J., and Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3), 302-319. doi: 10.1016/j.cognition.2013.02.013

Sonderegger, M., & Yu, A. (2010). A rational account of perceptual compensation for coarticulation. In S. Ohlsson & R. Catrambone (Eds.), Proceedings of the *32nd annual conference of the Cognitive Science Society* (pp. 375–380). Austin, TX: Cognitive Science Society.

Spivey-Knowlton, M. J., Trueswell, J. C., & Tanenhaus, M. K. (1993). Context effects in syntactic ambiguity resolution: Discourse and semantic influences in parsing reduced relative clauses. *Canadian Journal of Experimental Psychology*, 47(2), 276-309. doi: 10.1037/h0078826

Stevens, K. N., & Klatt, D. H. (1974). Role of formant transitions in the voiced-voiceless distinction for stops. *Journal of the Acoustical Society of America*, 55, 653–659.

Strand, E., & Johnson, K. (1996), October. Gradient and visual speaker normalization in the perception of fricatives. Natural language processing and speech technology: Results of the 3rd KONENS conference, ed. by D. Gibbon. Bielfeld: Germany.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, 331, 1279-1285. doi: 10.1126/science.1192788

Theodore, R. M., & Miller, J. L. (2010). Characteristics of listener sensitivity to talker-specific phonetic detail. *Journal of the Acoustical Society of America*, 128(4), 2090-2099. doi: 10.1121/1.3467771

Toscano, J. C. & McMurray, B. (2010). Cue integration with categories: Weighting acoustic cues in speech using unsupervised learning and distributional statistics. *Cognitive Science* 34(3), 434-464. doi: 10.1111/j.1551-6709.2009.01077.x

Trude, A. M., and Brown-Schmidt, S. (2012). Talker-specific perceptual adaptation during online speech perception. *Language and Cognitive Processes*, 27(7-8). 979-1001. doi: 10.1080/01690965.2011.597153

Trueswell, J. C., Tanenhaus, M. K., and Kello, C. (1993). Verb-specific constraints in sentence processing: Separating effects of lexical preference from garden-paths. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 19(3), 528-553. doi: 10.1037/0278-7393.19.3.528

Vallabha, G. K., McClelland, J. L., Pons, F., Werker, J. F., & Amano, S. (2007). Unsupervised learning of vowel categories from infant-directed speech. *Proceedings of the National Academy of Sciences of the United States of America*, 104(33), 13273–8. doi: 10.1073/pnas.0705369104

Vroomen, J., van Linden, S., de Gelder, B., & Bertelson, P. (2007). Visual recalibration and selective adaptation in auditory-visual speech perception: Contrasting build-up courses. *Neuropsychologia*, 45, 572-577. doi: 10.1016/j.neuropsychologia.2006.01.031

Walker, A., & Hay, J. (2011). Congruence between 'word age' and 'voice age' facilitates lexical access. *Laboratory Phonology*, 2(1). 219-237. doi: 10.1515/labphon.2011.007

Weatherholtz, K., & Jaeger, T. F. (to appear). Speech perception and generalization across talkers` and accents. *Oxford Research Encyclopedia in Linguistics*.

Xie, X., Theodore, M., & Myers, E. B. (under review). More than a boundary shift: Perceptual adaptation to foreign-accented speech reshapes the internal structure of phonetic categories.

Yildirim, I., Degen, J., Tanenhaus, M. K., & Jaeger, T. F. (2016). Talker-specificity and adaptation in quantifier interpretation. *Journal of Memory and Language*. 87, 128-143. doi: 10.1016/j.jml.2015.08.003