A Bayesian belief updating model of phonetic recalibration and selective adaptation

Dave Kleinschmidt[1] and **T. Florian Jaeger**[1,2]
Departments of [1]Brain and Cognitive Sciences and [2]Computer Science
University of Rochester
Rochester, NY, USA
{dkleinschmidt,fjaeger}@bcs.rochester.edu

Abstract

The mapping from phonetic categories to acoustic cue values is highly flexible, and adapts rapidly in response to exposure. There is currently, however, no theoretical framework which captures the range of this adaptation. We develop a novel approach to modeling phonetic adaptation via a belief-updating model, and demonstrate that this model naturally unifies two adaptation phenomena traditionally considered to be distinct.

1 Introduction

In order to understand speech, people map a continuous, acoustic signal onto discrete, linguistic categories, such as words. Despite a long history of research, no invariant mapping from acoustic features to underlying linguistic units has yet been found. Some of this lack of invariance is due to random factors, such as errors in production and perception, but much is due to systematic factors, such as differences between speakers, dialects/accents, and speech conditions.

The human speech perception system appears to deal with the lack of invariance in two ways: by storing separate, speaker-, group-, or context-specific representations of the same categories (Goldinger, 1998), and by rapidly adapting phonetic categories to acoustic input. Even though a person's inventory of native language phonetic categories is generally fixed from an early age (Werker and Tees, 1984), the mapping between these categories and their acoustic realizations is flexible. Listeners adapt rapidly to foreign-accented speech (Bradlow and Bent, 2008) and acoustically distorted speech (Davis et al., 2005), showing increased comprehension after little exposure. Such adaptation results in temporary and perhaps speaker-specific changes in phonetic categorization (Norris et al., 2003; Vroomen et al., 2007; Kraljic and Samuel, 2007).

To our knowledge, there is no theoretical framework which explains the range and specific patterns of adaptation of phonetic categories. In this paper, we propose a novel framework for understanding phonetic category adaptation—rational belief updating—and develop a computational model within this framework which straightforwardly explains two types of phonetic category adaptation which are traditionally considered to be separate.

While phonetic category adaptation has not thus far been described in this way, it nevertheless shows many hallmarks of rational inference under uncertainty (Jacobs and Kruschke, 2010). When there is another possible explanation for strange pronunciations (e.g. the speaker has a pen in her mouth), listeners do not show any adaptation (Kraljic et al., 2008). Listeners are more willing to generalize features of a foreign accent to new talkers if they were exposed to multiple talkers initially, rather than a single talker (Bradlow and Bent, 2008). Listeners also show rational patterns of generalizations of perceptual learning for specific phonetic contrasts, generalizing to new speakers only when the adapted phonetic categories of the old and new speakers share similar acoustic cue values (Kraljic and Samuel, 2007).

While it is not conclusive, the available evidence suggests that listeners update their beliefs about pho-

Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics, pages 10–19,
Portland, Oregon, June 2011.

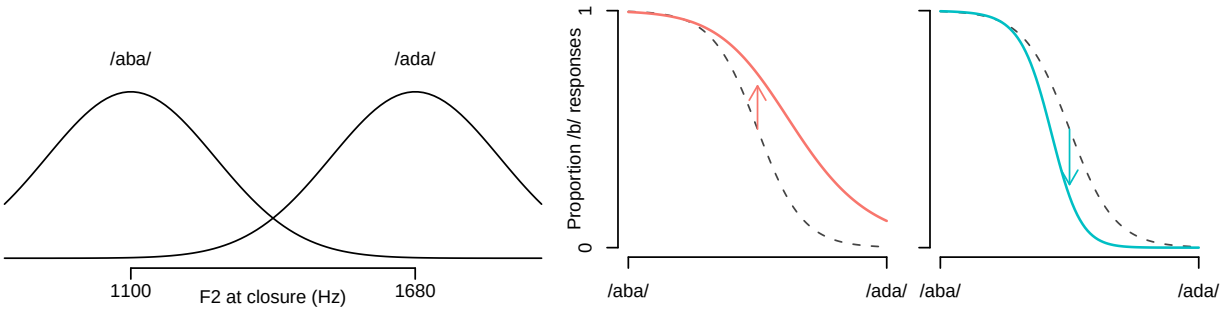

Figure 1: Left: approximate distribution of acoustic cue values for /aba/ and /ada/ stimuli from Vroomen et al. (2007). Right: exposure to acoustically ambiguous /aba/ tokens results in recalibration of the /aba/ category, with the classification boundary shifting towards /ada/ (center-right), while exposure to unambiguous /aba/ tokens results in selective adaptation of the /aba/ category, where the classification boundary shifts towards /aba/ (far right).

netic categories based on experience in a rational way. We propose that Bayesian belief updating can provide a principled computational framework for understanding rapid adaptation of phonetic categories as optimal inference under uncertainty. Such a framework has the appeal of being successfully applied in other domains (Brenner et al., 2000; Fine et al., 2010). In addition, rational models have also been used within the domain of speech perception to model acquisition of phonetic categories (Vallabha et al., 2007; Feldman et al., 2009a; McMurray et al., 2009), the perceptual magnet effect (Feldman et al., 2009b), and how various cues to the same phonetic contrast can be combined (Toscano and McMurray, 2010).

2 The Phenomena: Perceptual recalibration and selective adaptation

The flexibility of phonetic categories has been demonstrated through studies which manipulate the distribution of acoustic cues associated with a particular category. These studies take advantage of the natural variability of acoustic cues. Take, for example, the consonants /b/ and /d/. These two consonants can be distinguished largely on the basis of the trajectory of the second formant before and after closure (Iskarous et al., 2010). Like all acoustic-phonetic cues, there is natural variability in the F2 locus for productions of each category (depicted schematically in Figure 1, left). Listeners react to subtle changes in the distributions of acoustic cues, and adjust their phonetic categories for a variety of contrasts and manipulations (Kraljic and Samuel, 2006). In this paper, we model the effects of the two most common types of manipulation studied thus far, which produce opposite changes in phonetic classification.

The first of these is repeated exposure to acoustically ambiguous tokens, which results in a change in classification termed "perceptual learning" (Norris et al., 2003) or "perceptual recalibration" (Bertelson et al., 2003) in which the initially-ambiguous token becomes an accepted example of one phonetic category. Such ambiguous cue values are not uncommon because of the natural variability in normal speech. It is thus possible to generate a synthetic production /?/ which is acoustically intermediate between /b/ and /d/, and which is phonetically ambiguous in the absence of other cues but nevertheless sounds like a plausible production. When paired with another cue which implies /b/, subjects reliably classify /?/ as /b/. Disambiguating information could be provided by a video of a talker producing /b/ (Vroomen et al., 2007), or a word such as *a?out*, where a /b/ has been replaced with /?/ (Norris et al., 2003). When /?/ is repeatedly paired in this way with information biasing a /b/ interpretation, subjects begin to interpret /?/ as /b/ in general, classifying more items on a /b/-to-/d/ continuum as /b/ (Figure 1, center-right, red curve).

A second manipulation is repeated exposure to the same, acoustically unambiguous token. Repeated exposure to /b/ causes "selective adaptation" of this category, where listeners are less likely to

classify items as /b/, indicated by a shift in the /b/-/d/ classification boundary towards /b/ (Figure 1, far-right).

Traditionally, recalibration and selective adaptation have been analyzed as separate processes, driven by separate underlying mechanisms (Vroomen et al., 2004), since they arise under different circumstances and produce opposite effects on classification. They also show different time courses. Vroomen et al. (2007) found that, on the one hand, strong recalibration effects occur after just a few exposures to ambiguous tokens, but fade with further exposure (Figure 3, upper curve). On the other, selective adaptation is present after a few exposures to unambiguous tokens, but grows steadily stronger with further exposure (Figure 3, lower curve).

We will show that these two superficially different adaptation phenomena are actually closely related, and will provide a unified account by appealing to principles of Bayesian belief updating. These principles are used to construct two models. The first, a unimodal model, treats phonetic categories as distributions over acoustic cue dimensions. The second, a multimodal model, treats phonetic categories as distributions over *phonetic* cue dimensions, which integrate information from both audio and visual cues. Both models capture the general effect directions of selective adaptation and recalibration, but only the multimodal model captures their distinct time courses.

The next section provides a high-level descriptions of these models, and how they might describe the selective adaptation and recalibration data of Vroomen et al. (2007). Section 4 describes this data and the methods used to collect it in more details. Section 5 describes the general modeling framework, how it was fit to the data, and the results, and Section 6 describes the multimodal model and its fit to the data.

3 Phonetic category adaptation via belief updating

In our proposed framework, the listener's classification behavior can be viewed as arising from their beliefs about the distribution of acoustic cues for each phonetic category. Specifically, as we will develop more rigorously below, the probability of classifying a given token x (which is the value of either an acoustic cue or a multimodal, phonetic cue) as /b/ is proportional to the relative likelihood of the cue value x arising from /b/ (relative to the overall likelihood of observing tokens like x, regardless of category). Thus, changes in the listener's beliefs about the distribution of cue values of category /b/ will result in changes in their willingness to classify tokens as /b/.

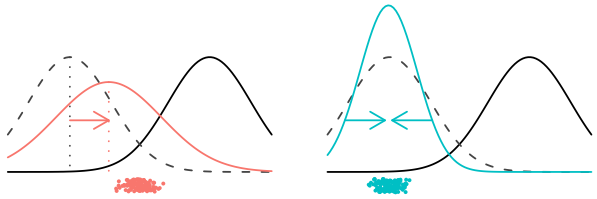

Figure 2: An incremental belief-updating model for phonetic recalibration and selective adaptation. These distributions correspond to the classification functions in Figure 1. Left: ambiguous stimuli labeled as /b/ cause a shift of the /b/ category towards those stimuli. Right: repeated unambiguous stimuli correspond to a narrower distribution than expected.

A belief-updating model accounts for recalibration and selective adaptation in the following way. When, on the one hand, a listener encounters many tokens that they consider to be /b/ but which are all acoustically intermediate between /b/ and /d/, they will change their beliefs about the distribution of /b/, shifting it to better align with these ambiguous cue values (Figure 2, left). This results in increased categorization of items on a /b/-to-/d/ continuum as /b/, since the range on the continuum over which the likelihood associated with /b/ is higher than that of /d/ is extended.

On the other hand, when a listener encounters repeated, tightly-clustered and highly prototypical /b/ productions, they update their beliefs about the distribution of /b/ to reflect that /b/ productions are more precise than they previously believed (Figure 2, right). They consequently assign lower likelihood to intermediate, ambiguous cue values for /b/, causing them to classify fewer /b/-/d/ continuum items as /b/.

Modeling the time course of selective adaptation

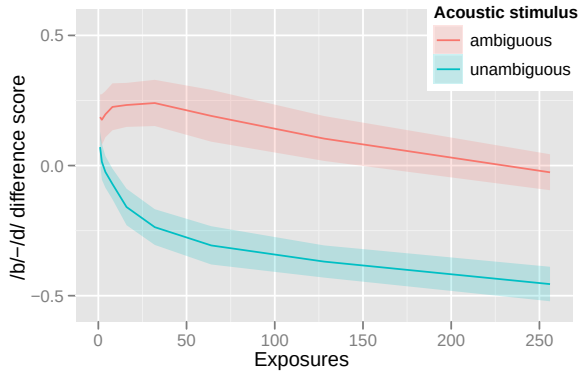

Figure 3: The results of Vroomen et al. (2007), showing the build-up time course of selective adaptation (as a function of unambiguous exposure trials) and recalibration (as a function of ambiguous exposure trials).

is straightforward: the more observations are made, the narrower the distribution becomes, and the more the classification boundary shifts towards the adapting category. However, modeling the time course of recalibration, as measured by Vroomen et al. (2007), is more complicated. Recalibration comes on quickly, but fades gradually with many exposures (Figure 3). As discussed below in Section 5.3, the unimodal model cannot account for this pattern, because it consideres the acoustically-similar exposure and test stimuli the same. The multimodal model, by integrating audio and visual cues to form the adapting percept, dissociates the adapting stimulus from the test stimuli and does not suffer from this problem. It is thus in principle capable of reproducing the empirical time course of recalibration observed by Vroomen et al. (2007). In practice, this model does indeed provide a good qualitative fit to human data, as discussed in Section 6.

4 Behavioral data: Vroomen et al. (2007)

Vroomen et al. (2007) investigated the time course of adaptation to audio-visual speech stimuli. In each block, subjects were repeatedly exposed to a single type of stimulus. The visual stimulus was either /aba/ or /ada/, and the audio stimulus was either an unambiguous match of the visual stimulus or was an ambiguous production. Throughout exposure, subjects were tested with unimodal acoustic test stimuli in order to measure the effect of exposure thus far.

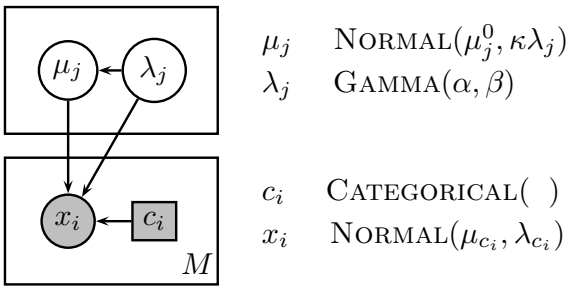

Figure 4: Graphical model for the mixture of Gaussians with normal-gamma prior model. See text for description.

The overall effect of exposure to unambiguous stimuli was computed by comparing classification between unambiguous-/b/ and unambiguous-/d/ exposure, and likewise for the effect of exposure to ambiguous stimuli.

The acoustic stimuli used in exposure and test were drawn from a nine-item continuum (denoted $x = 1, \ldots, 9$) from /aba/ to /ada/, formed by manipulating the second formant frequency before and after the stop consonant (Vroomen et al., 2004). The most /aba/-like item $x = 1$ was synthesized using the formant values from a normal /aba/ production, and the most /ada/-like item $x = 9$ was derived from an /ada/ production. The maximally ambiguous item was determined for each subject via a labeling function (percent-/aba/ classification for each token) derived from pre-test classification data (98 trials from across the entire continuum). All subjects' maximally ambiguous tokens were one of $x = 4, 5$ or 6.

Each exposure block consisted of 256 repetitions of the bimodal exposure stimulus. After 1, 2, 4, 8, 16, 32, 64, 128, and 256 exposure trials subjects completed a test block, of six classification trials. They were asked to classify as /aba/ or /ada/ the three most ambiguous stimuli from the continuum (the most ambiguous stimulus and the two neighboring stimuli) twice each. For each ambiguity condition, the aggregate effect of exposure across categories was a difference score, calculated by subtracting the percent /aba/-classification after /d/-exposure from the percent after /b/-exposure. This /b/-/d/ difference score, as a function of cumulative exposure trials, is plotted in Figure 3.

5 The unimodal model

We implemented an incremental belief-updating model using a mixture of Gaussians as the underlying model of phonetic categories (Figure 4), where each phonetic category $j = 1 \ldots M$ corresponds to a normal distribution over percepts x with mean μ_j and precision (inverse-variance) λ_j (e.g. Figure 1, left).

$$p(x_i \mid c_i) = \mathcal{N}(\mu_{c_i}, \lambda_{c_i}) \quad (1)$$

The listener's beliefs about phonetic categories are captured by additionally assigning probability distributions to the means μ_j and precisions λ_j of each phonetic category. The prior distribution $p(\mu_j, \lambda_j)$ represents the listener's beliefs before exposure to the experimental stimuli, and the posterior $p(\mu_j, \lambda_j \mid X)$ captures the listener's beliefs after exposure to stimuli X from category j. These two distributions are related via Bayes' Rule:

$$p(\mu_j, \lambda_j \mid X) \propto p(X \mid \mu_j, \lambda_j) p(\mu_j, \lambda_j) \quad (2)$$

In order to quantitatively evaluate such a model, the form of the prior distributions needs to be specified. A natural prior to use in this case is known as a Normal-Gamma prior.[1] This prior factorizes the joint prior into

$$\begin{aligned} p(\mu_j, \lambda_j) &= p(\mu_j \mid \lambda_j) p(\lambda_j) \\ p(\mu_j \mid \lambda_j) &= \mathcal{N}(\mu_j^0, \kappa\lambda_j) \\ p(\lambda_j) &= \mathcal{G}(\alpha, \beta) \end{aligned}$$

where $\mathcal{N}(\mu_j^0, \kappa\lambda_j)$ is a Normal distribution with mean μ_j^0 and precision $\kappa\lambda_j$, and $\mathcal{G}(\alpha, \beta)$ is a Gamma distribution with shape α and rate β (Figure 4).

5.1 Identifying individual subjects' prior beliefs

In order to pick the most ambiguous token for each subject, Vroomen et al. (2007) collected calibration data from their subjects, which consisted of 98 two-alternative forced choice trials on acoustic tokens spanning the entire /aba/-to-/ada/ continuum. As revealed by this pre-test data, each subject's phonetic categories are different, and so we chose to estimate the prior beliefs about the nature of the exposure categories on a subject-by-subject basis. We fit each subject's classification function using logistic regression. The logistic function is closely related to the distribution over category labels given observations in a mixture of Gaussians model. Specifically, when there are only two categories (as in our case), the probability that an observation at x will be labeled c_1 is[2]

$$p(c_1 \mid x) = \frac{p(x \mid c_1)p(c_1)}{p(x \mid c_1)p(c_1) + p(x \mid c_2)p(c_2)} \quad (3)$$

Further assuming that the categories have equal precision λ and equal prior probability $p(c_1) = p(c_2) = 0.5$[3], this reduces to a logistic function of the form $p(c_1 \mid x) = (1 + \exp(-gx + b))^{-1}$, where

$$g = (\mu_1 - \mu_2)\lambda \quad \text{and} \quad b = (\mu_1^2 - \mu_2^2)\lambda$$

Even when b and g can be estimated from the subject's pre-test data, one additional degree of freedom needs to be fixed, and we chose to fix the distance between the means, $\mu_1 - \mu_2$. Given these values, the values for $(\mu_1 + \mu_2)/2$ (the middle of the subject's continuum) and λ can be calculated using

$$\frac{\mu_1 + \mu_2}{2} = \frac{b}{g} \quad \text{and} \quad \lambda = \frac{g}{\mu_1 - \mu_2} \quad (4)$$

We chose to use $\mu_1 - \mu_2 = 8$, the length of the acoustic continuum, which stretches from $x = 1$ (derived from a natural /aba/) to $x = 9$ (from a natural /ada/). This is roughly equivalent to assuming that all subjects would accept these tokens as good productions of /aba/ and /ada/, which indeed they do (Vroomen et al., 2004).

So far, we have accounted for the expected values of category means and precisions. The strength of these prior beliefs, however, has yet to be specified, and unfortunately there is no way to estimate this based on the pre-test data of Vroomen et al. (2007). The model parameters corresponding to the

[1] It is natural in that the Normal-Gamma distribution is the conjugate prior for a Gaussian distribution where there is some uncertainty about both the mean and the precision. Using the conjugate prior ensures that the posterior distribution has the same form as the prior.

[2] Here we are abusing notation a bit by using c_1 as a shorthand for $c = 1$.

[3] This assumption is not strictly necessary, but for this preliminary model we chose to make it in order to keep the model as simple as possible.

subject's confidence in their prior beliefs are κ and α for the means and variances, respectively. Given the specific form of the prior we use here, these two parameters are closely related to the number of observations that are required to modify the subject's belief about a phonetic category (Murphy, 2007).

5.2 Model fitting

In order to evaluate the performance of this model relative to human subjects, four simulations were run per subject, corresponding to the four conditions used by Vroomen et al. (2007): ambiguous /d/ and /b/, and unambiguous /d/ and /b/. For each subject, the hyper-parameters $(\mu_j^0, \kappa, \alpha, \beta)$ were set according to the methods described above: values were chosen for the free parameters α and κ, and β and μ_j^0 were set based on the subject's pre-test data.

To model the effect of n exposure trials in a given condition, the stimuli used by Vroomen et al. (2007) were input into the model in the following way. For ambiguous blocks, the observations X were n repetitions of that subject's most ambiguous token, and for unambiguous blocks they were n repetitions of the $x = 1$ for /b/ or $x = 9$ for /d/. For /b/ exposure blocks, the category labels C were set to 1, and for /d/ they were set to 2, corresponding to the disambiguating effect of the visual cues.

For each subject, condition, and number of exposures, the posterior distribution over category means and precisions $p(\mu_j, \lambda_j \mid X, C)$ was sampled using numerical MCMC techniques.[4]

To compare the simulation results with the test data of Vroomen et al. (2007), it was necessary to find the classification function, $p(c_{\text{test}} = 1 \mid x_{\text{test}}, X)$, which is the probability that acoustic test stimulus x_{test} will be categorized as /b/ ($c_{\text{test}} = 1$) given the training data X. Based on (3), it suffices to find the predictive distributions

$$p(x_{\text{test}} \mid c_{\text{test}} = 1, X) = \iint p(x_{\text{test}} \mid \mu_1, \lambda_1) p(\mu_1, \lambda_1 \mid X) \mathrm{d}\mu_1 \mathrm{d}\lambda_1$$

and, analogously, $p(x_{\text{test}} \mid c_{\text{test}} = 2, X)$. These integrals can be approximated numerically, by averaging over the individual likelihoods corresponding to each individual pair of means and variances drawn from the posterior $p(\mu_j, \lambda_j \mid X)$.

[4] Specifically, 1 000 samples for each parameter were obtained after burn-in using JAGS, an open-source implementation of the BUGS language for Gibbs sampling of graphical models: `https://sourceforge.net/projects/mcmc-jags`

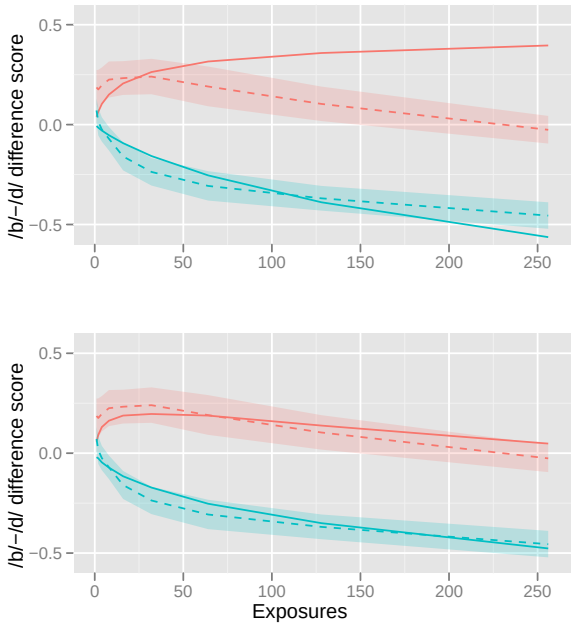

Figure 5: Overall fit of the acoustic-only (top, $R^2 = 0.14$) and bimodal model (bottom $R^2 = 0.67$). Solid lines correspond to the best fit averaged over subjects, and dashed lines correspond to empirical difference scores, with shaded regions corresponding to the 95% confidence interval on the empirical subject means.

Once this labeling function is obtained, the dependent measure used by Vroomen et al. (2007)—average percentage categorized as /b/—can be calculated, by averaging the value of $p(c_{\text{test}} = 1 \mid x_{\text{test}}, X)$ for the test stimuli x_{test} used by Vroomen et al. (2007). These were the subject's maximally ambiguous stimulus ($x = 4, 5$ or 6, depending on the subject), and its two neighbors on the continuum. The difference score used by Vroomen et al. (2007) was computed by subtracting the average probability of /b/ classification after /b/ ($c = 1$) exposure from the probability of /b/ classification after /d/ ($c = 2$) exposure. The best fitting confidence parameters α and κ were those which minimized mean squared error between the empirical and model difference scores.

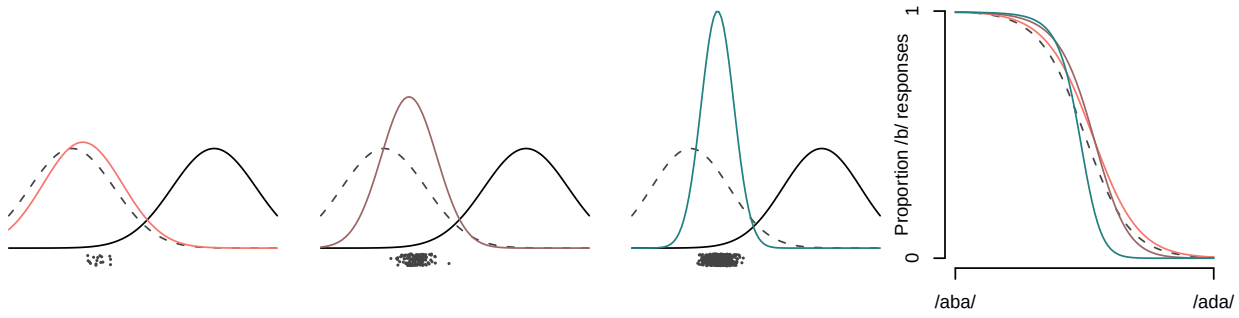

Figure 6: When audio and visual cues are integrated before categorization, a small number of ambiguous tokens still produces a shift in the category mean, and thus recalibration (left, bright red). However, a large number of ambiguous tokens produces both a shift of the category mean and an increase in precision (center-right, dark blue). If the audio-visual percept is located away from the maximally ambiguous middle region of the continuum, this can result in an extinction of the initial recalibration effect with increasing exposure (far right).

5.3 Results

Figure 5, top panel shows the results of the unimodal model. While this model clearly captures the direction of the effects caused by ambiguous and unambiguous exposure, it fails to account for a significant qualitative feature of the human data: the rise and then fall of the recalibration effect (red line).

The reason for this is that the audio component of the audio-visual exposure stimuli is identical to the maximally ambiguous (audio-only) test stimulus. Under this model, the probability with which a stimulus is classified as /b/ is proportional to the likelihood assigned to that cue value by category /b/, relative to the total likelihood assigned by /b/ and /d/. In addition, under rational belief updating the likelihood assigned to the exposure stimulus' cue value will always increase with more exposure. In the unimodal model the cue dimension is only auditory (with the visual information in the exposure stimuli only being used to assign category labels), and so to the unimodal model the ambiguous exposure stimuli and the ambiguous test stimuli are exactly the same. Thus, the probability that the test stimuli will be categorized as the exposure category increases monotonically with further exposure.

6 The multimodal model

The unimodal model assumes that the cue dimensions which phonetic categories are defined over are *acoustic*, incorporating information from other modalities only indirectly. This assumption is almost certainly wrong, based on work on audio-visual speech, which shows strong and pervasive cross-modal interactions (McGurk and MacDonald, 1976; Bejjanki et al., 2011). Indeed, Bertelson et al. (2003) report strong effects of the visual cue used by Vroomen et al. (2007): subjects were at chance in discriminating acoustically ambiguous versus unambiguous bimodal tokens when the visual cue matched.

The multimodal model replaces the acoustic percept x in the unimodal model with a phonetic percept which integrates information from audio and visual cues. Under reasonably general assumptions, information from auditory and visual cues to the same phonetic dimension can be optimally combined by a simple weighted sum $x = w_a x_a + w_v x_v$, where the weights w_a and w_v sum to 1 and are proportional to the reliability of the auditory and visual cues (Ernst and Banks, 2002; Knill and Saunders, 2003; Jacobs, 2002; Toscano and McMurray, 2010).

Such optimal linear cue-combination can be incorporated into our model in an approximate way by replacing x with a weighted sum of the continuum values for the auditory and visual tokens $x = wx_a + (1-w)x_v$. In the unambiguous conditions, there is no mismatch between these values ($x_a, x_v = 1$ for /aba/ trials and 9 for /ada/ trials), and behavior is the same. In the ambiguous trials, however, the combination of visual and auditory cues creates a McGurk illusion, and pulls the observed stimulus—now located on a phonetic /aba/-/ada/ continuum rather than an acoustic one—away

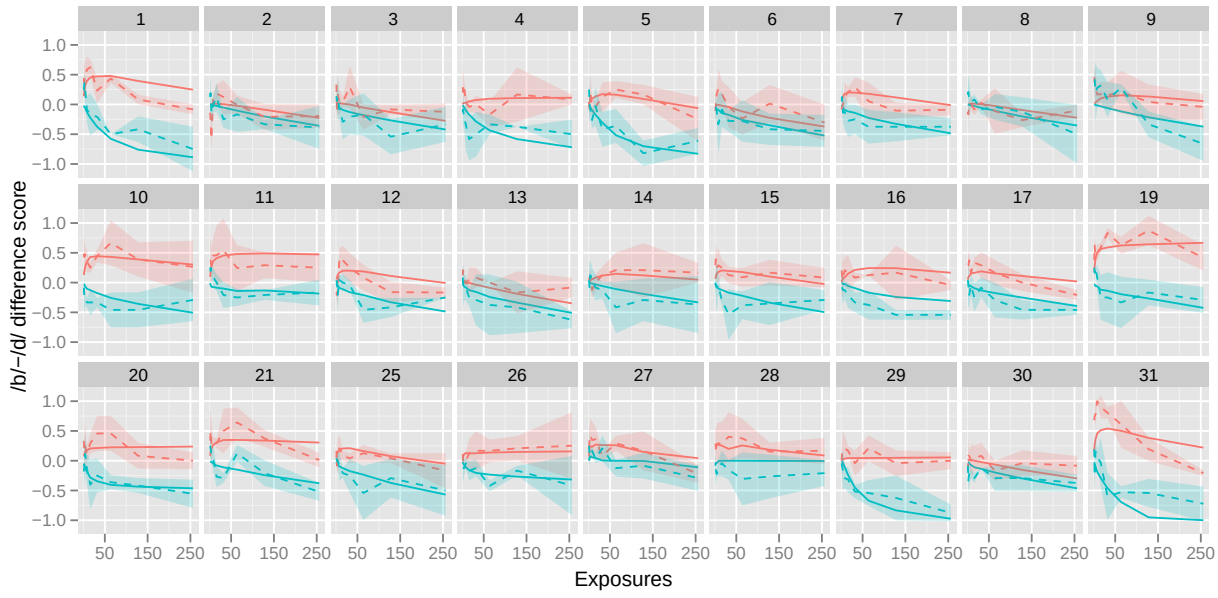

Figure 7: Best model fit for each individual subject. Dashed lines are empirical difference scores (shaded regions are 95% confidence intervals) and solid lines are the best-fitting model for that subject. Mean $R^2 = 0.57$, SE$= 0.04$.

from the maximally ambiguous test stimuli, which are still located at the middle of the continuum, being audio-only. This allows recalibration to dominate early, as the mean of the adapted category moves towards the adapting percept, but be reversed later, as the precision increases with further exposure percepts, all tightly clustered around the new, intermediate mean (Figure 6).

To be optimal, w must be the relative reliability (precision) of audio cues relative to visual cues, but in this preliminary model it is treated as a free parameter, between 0 and 1, and fit to each subject's test data individually, in the same way as the confidence parameters α and κ.

The best fitting models' predictions are shown averaged across subjects in Figure 5 (bottom panel). Unlike the unimodal model, the multimodal model clearly captures the initial rise and later fall of recalibration for ambiguous stimuli, and captures a fair amount of the variation between subjects (Figure 7).

7 Discussion

The Bayesian belief updating model developed in this paper, which takes into account cross-modal cue integration, provides a good qualitative fit to both the overall direction and detailed time-course of two very different types of adaptation of phonetic categories, recalibration and selective adaptation, as studied by Vroomen et al. (2007). This constitutes a first step towards a novel theoretical framework for understanding the flexibility that characterizes the mapping between phonetic categories to acoustic (and other) cues. There is a large number of models which adhere to the basic principles outlined here, and we have investigated only two of the simplest ones in order to show that, firstly, selective adaptation and recalibration can be considered the product of the same underlying inferential process, and secondly, this process likely occurs at the level of multimodal phonetic percepts.

One of the most striking findings from this work, which space precludes discussing in depth, is that all subjects' data is fit best when the strength of the prior beliefs is quite low, corresponding to a few hundred or thousand prior examples, which is many orders of magnitude less than the number of /b/s and /d/s a normal adult has encountered in their life. Why should this number be so low? The answer lies in the fact that phonetic adaptation is often extremely specific, at the level of a single speaker or situation. In the future, we plan to model these patterns of specificity and generalization (Kraljic and

Samuel, 2007; Kraljic and Samuel, 2006) via hierarchical extensions of the current model, with connected mixtures of Gaussians for phonetic categories that vary in predictable ways between groups of speakers.

Besides being a principled, mathematical framework, Bayesian belief updating and the broader framework of rational inference under uncertainty also provides a good framework for understanding how and why multiple cues are combined in phonetic categorization (Toscano and McMurray, 2010; Jacobs, 2002). Finally, this approach is similar in spirit and in its mathematical formalisms to models which treat the acquisition of phonetic categories as statistical inference, where the number of categories needs to be inferred, as well as the means and precisions of those categories (Vallabha et al., 2007; Feldman et al., 2009a). It is also similar to recent work on syntactic adaptation (Fine et al., 2010), and thus constitutes a central part of an emerging paradigm for understanding language as inference and learning under uncertain conditions.

Acknowledgements

We would like to thank Jean Vroomen for generously making the raw data from Vroomen et al. (2007) available.

This work was partially funded by NSF Grant BCS-0844472 and an Alfred P. Sloan Fellowship to TFJ.

References

Vikranth Rao Bejjanki, Meghan A Clayards, David C Knill, and Richard N Aslin. 2011. Cue Integration in Categorical Tasks : Insights from Audio-Visual Speech Perception. *PLoS ONE*, in press.

Paul Bertelson, Jean Vroomen, and Béatrice de Gelder. 2003. Visual recalibration of auditory speech identification: a McGurk aftereffect. *Psychological Science*, 14(6):592–597, November.

Ann R Bradlow and Tessa Bent. 2008. Perceptual adaptation to non-native speech. *Cognition*, 106(2):707–29, February.

Naama Brenner, William Bialek, and Rob de Ruyter Van Steveninck. 2000. Adaptive Rescaling Maximizes Information Transmission. *Neuron*, 26(3):695–702, June.

Matthew H Davis, Ingrid S Johnsrude, Alexis Hervais-Adelman, Karen Taylor, and Carolyn McGettigan. 2005. Lexical information drives perceptual learning of distorted speech: evidence from the comprehension of noise-vocoded sentences. *Journal of experimental psychology. General*, 134(2):222–41, May.

Marc O Ernst and Martin S Banks. 2002. Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, 415(6870):429–33.

Naomi H Feldman, Thomas L Griffiths, and James L Morgan. 2009a. Learning phonetic categories by learning a lexicon. *Proceedings of the 31st Annual Conference of the Cognitive Science Society*, pages 2208–2213.

Naomi H Feldman, Thomas L Griffiths, and James L Morgan. 2009b. The influence of categories on perception: explaining the perceptual magnet effect as optimal statistical inference. *Psychological review*, 116(4):752–82, October.

Alex B Fine, Ting Qian, T Florian Jaeger, and Robert A Jacobs. 2010. Is there syntactic adaptation in language comprehension? In *ACL Workshop on Cognitive Modeling and Computational Linguistics*, pages 18–26.

Stephen D Goldinger. 1998. Echoes of echoes? An episodic theory of lexical access. *Psychological review*, 105(2):251–79, April.

Khalil Iskarous, Carol A Fowler, and D H Whalen. 2010. Locus equations are an acoustic expression of articulator synergy. *The Journal of the Acoustical Society of America*, 128(4):2021–32, October.

Robert A Jacobs and John K Kruschke. 2010. Bayesian learning theory applied to human cognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, pages n/a–n/a, May.

Robert A Jacobs. 2002. What determines visual cue reliability? *Trends in cognitive sciences*, 6(8):345–350, August.

David C Knill and Jeffrey A Saunders. 2003. Do humans optimally integrate stereo and texture information for judgments of surface slant? *Vision Research*, 43(24):2539–2558, November.

Tanya Kraljic and Arthur G Samuel. 2006. Generalization in perceptual learning for speech. *Psychonomic bulletin & review*, 13(2):262–8, April.

Tanya Kraljic and Arthur G Samuel. 2007. Perceptual adjustments to multiple speakers. *Journal of Memory and Language*, 56(1):1–15, January.

Tanya Kraljic, Arthur G Samuel, and Susan E Brennan. 2008. First impressions and last resorts: how listeners adjust to speaker variability. *Psychological science : a journal of the American Psychological Society / APS*, 19(4):332–8, April.

Harry McGurk and John MacDonald. 1976. Hearing lips and seeing voices. *Nature*, 264(5588):746–748.

Bob McMurray, Richard N Aslin, and Joseph C Toscano. 2009. Statistical learning of phonetic categories: insights from a computational approach. *Developmental Science*, 12(3):369–78, April.

Kevin P Murphy. 2007. Conjugate Bayesian analysis of the Gaussian distribution. *Technical report, University of British Columbia*.

Dennis Norris, James M McQueen, and Anne Cutler. 2003. Perceptual learning in speech. *Cognitive Psychology*, 47(2):204–238, September.

Joseph C Toscano and Bob McMurray. 2010. Cue integration with categories: Weighting acoustic cues in speech using unsupervised learning and distributional statistics. *Cognitive science*, 34(3):434–464, April.

Gautam K Vallabha, James L McClelland, Ferran Pons, Janet F Werker, and Shigeaki Amano. 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, August.

Jean Vroomen, Sabine van Linden, Mirjam Keetels, Béatrice de Gelder, and Paul Bertelson. 2004. Selective adaptation and recalibration of auditory speech by lipread information: dissipation. *Speech Communication*, 44(1-4):55–61, October.

Jean Vroomen, Sabine van Linden, Béatrice de Gelder, and Paul Bertelson. 2007. Visual recalibration and selective adaptation in auditory-visual speech perception: Contrasting build-up courses. *Neuropsychologia*, 45(3):572–7, February.

Janet F Werker and Richard C Tees. 1984. Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development*, 7(1):49–63, January.