A Distributional Account of Covariance Effects and Talker Adaptation in Infant and Adult Phonetic Category Recognition

Bevan K. Jones (Bevan_Jones@Brown.edu)
Department of Cognitive and Linguistic Sciences
Brown University
Providence, RI 02912, USA

Abstract

Both infants and adults are sensitive to the non-linguistic features of speech, and this sensitivity impacts speech sound categorization, but with somewhat different effects. While both infants and adults sometimes confuse the non-linguistic for the linguistic and are susceptible to categorization problems when the two covary, adults, on the other hand, are often able to exploit non-linguistic features to improve perceptual categorization. We present a Bayesian account of both adult and infant behavior, arguing that differing levels in linguistic maturity correspond to different models of linguistic structure. The infant's task is one of structure learning, adults, on the other hand, are estimating parameters for an already established structure.

Keywords: Speech perception; distributional learning; language acquisition; Bayesian models.

Introduction

Talker variability is a fundamental challenge in speech perception. The same phonetic category as uttered by two different talkers may seem quit different. At the same time, distinct categories produced by two different talkers may be acoustically quite similar (Dorman, Studdert-Kennedy, & Raphael, 1977; Magnuson & Nusbaum, 2007). Unsurprisingly, this variability poses a problem for infants as they acquire their language. In particular, studies have shown that infants are prone to confounding talker-specific characteristics with phonetic categories when the talker covaries with the category during learning (Houston & Jusczyk, 2000; Creel, Aslin, & Tanenhaus, 2008). For instance, when taught to recognize two different categories, one produced exclusively by a female speaker and the other by a male speaker, infants were unable later to identify those phones when spoken by the opposite sex. This suggests that learning not only involves acquiring information about the features of the exemplars of the category, but, more fundamentally, about which features relate to the categorization task at all.

Adults are not immune to talker variability either and can also be misled by talker differences (Kraljic, Brennan, & Samuel, 2008; McQueen, Norris, & Cutler, 2006), but the same studies also demonstrate that adults are able to adapt to the differences. In fact, speaker identity may even be exploited to improve recognition performance at times, as suggested by experiments with episodic memory. Goldinger (1996) showed that words spoken by one speaker can be more easily recognized when uttered by the same speaker even after significant time has elapsed. This suggests that not only do listeners note linguistically weighted cues but also indexical cues that might be used for talker identification.

While both infants and adults are faced with similar input and utilize statistical learning mechanisms, the nature of the problem they each face is quite different. Both face a categorization problem. Infants are still struggling to decide which dimensions in the high dimensional perceptual space are most relevant to the categorization task. Voice onset time, for instance, serves largely to distinguish the words “dime” and “time” since “d” is followed by a much shorter voicing delay than “t”. Other features such as fundamental frequency may serve an indexical function (aiding in distinguishing whether the talker is male or female, for instance) but are much less clearly related to the linguistic content in a language like English. Infants are engaged in a kind of feature selection, narrowing down the infinite set of possible features to just those that are most useful. Adults, on the other hand, have already determined which features are linguistic and which are not. However, far from simply discarding the non-linguistic information, adults may employ indexical features to track the talker, allowing them to adapt to the peculiarities of the individual's speech patterns.

We present a Bayesian account for both the infant and adult behavioral results. In the infant's case, the problem can be framed in terms of a model selection problem, a search through some space of models that relate the latent phonetic category to the observed features, both linguistic and non-linguistic. In the adult's case, talker adaptation is more of a problem of parameter estimation given an already learned model relating phonetic category, talker, and the observed linguistic and indexical features.

The models we present fall within the distributional learning paradigm. It is well known that speech sounds of all types tend to fall according to a Gaussian distribution (Peterson & Barney, 1951; Lisker & Abramson, 1964; Espy-Wilson, 1992). Furthermore, Maye, Werker, and Gerken (2002) show that bimodal distributions tend to prompt infants to identify two sounds where unimodal distributions lead to identification of a single category, suggesting that learners may rely to some extent on an assumption of something like a Gaussian distribution. Thus, learning can be characterized as a kind of parametric statistical search over unimodal or, in our case, Gaussian distributions.

We present an array of models to account for the different behaviors, arguing that not one, but several different models of the dependencies between features are required. Linguistic development is characterized under our assumption of multiple models as the selection of one model over another based on accumulated evidence. In the early days, when infants have little evidence of which model is likely to generalize, infants make decisions based on recent experience. Hence, covarying talker with phonetic category during training results in the infant's selecting a model that does not generalize to a more natural situation where talker and phonetic category do not covary. Similarly, we argue that adult talkers also shift between models depending on the available information. In the adult's case experience is not so acute an issue, but some features are not always present in the input, or are obscured by noise, and thus they must use an alternative model that does not depend on those features.

We argue for a fluid shifting between models over a single monolithic model. Shifts between qualitatively different models, as opposed to a gradual adjustment of a single model, accounts for how distinct situations result in different processes. Yet each model operates on the same basic principles of distributional learning, where even the shift between models may be accounted for within a Bayesian framework.

Model Definitions

Figure 1 presents the four different structural relationships we consider, slight variations but with important implications. At heart, they are all instances of a Gaussian mixture model which attempts to explain the linguistic feature x_i of the i^{th} sound by a distribution indexed by the sound's phonetic category c_i. The more complex models ($\mathcal{M}_3$ and $\mathcal{M}_4$) elaborate on the theme by introducing talker specific distributions over x_i, and introduce an additional latent variable t_i for each sound to represent talker identity. All the models assume exactly two phonetic categories, and the talker specific models in turn assume exactly two talkers, a restriction that is easily relaxed but does not interfere with our purpose: explaining the human behavior in certain psycholinguistic experiments.

In the case of models $\mathcal{M}_1$ and $\mathcal{M}_3$ each speech sound also bears an indexical feature y. The two models treat y quite differently, however. $\mathcal{M}_1$ assumes all features are linguistic, and therefore represents a direct dependency between c_i and y_i, paralleling the dependency between c_i and x_i. $\mathcal{M}_3$, however, distinguishes between linguistic and indexical features, and introduces a direct dependency between the indexical feature and the talker instead of the phonetic category. This change captures the notion that indexical features primarily serve to identify the talker, and only secondarily aid in recognition. This feature could be anything: fundamental frequency, or even an odd way of smacking ones lips at the end of each utterance. Since we are primarily interested in modeling phonetic category learning and not so much talker recognition, we treat this feature as a simple Bernoulli variable with a predefined parameter. That is, while the model learns the parameters for the distributions over x, y is determined by a prespecified Bernoulli parameter.

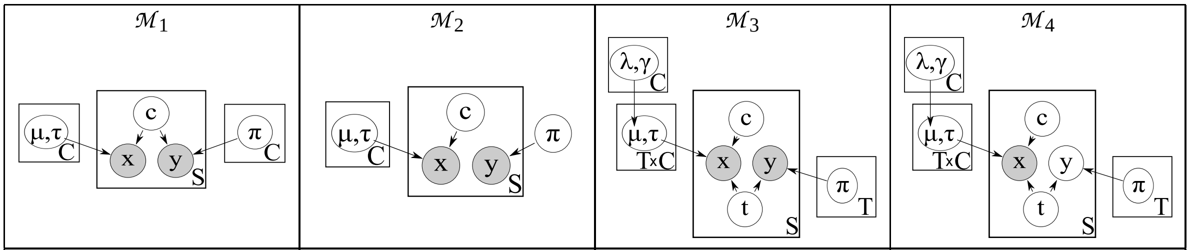
These models attempt to explain the phenomena observed in certain psycholinguistic experiments. Houston and Jusczyk (2000) demonstrated that 7.5 month olds were able to recognize words in a segmentation task when they were produced by a speaker of the same sex during test time as during training, but were unable to generalize across sexes. Singh (2008) demonstrates a similar sensitivity to other covariant non-linguistic features. Model $\mathcal{M}_1$ captures the behavior of infants in these situations, where all features are treated as linguistic. Since the model assumes all features are directly relevant to the categorization task, it will have a tendency to over fit when presented with data where talker and phonetic category accidentally covary (or are contrived to do so by an experimenter). Model $\mathcal{M}_2$, on the other hand, treats the indexical feature as independent, only modeling the dependency between x and c, and is more likely to generalize across speakers.

Models $\mathcal{M}_3$ and $\mathcal{M}_4$ introduce the ability to adapt to individual talkers by providing separate talker-specific distributions for the linguistic feature x. However, the individual talker-specific distributions for a particular phonetic category are related to each other by a distribution for the category common to all talkers. Thus, we introduce a hierarchical Gaussian distribution over linguistic features, capturing the notion that, although each talker may have his own peculiar way of producing a sound, sounds of the same category all tend to be similar across speakers. The hierarchical distribution allows for speech recognition even when faced with a completely unfamiliar talker, since the λ and γ parameters define a prior over talker specific categories, providing a mechanism of generalization from familiar talkers to novel talkers.

Goldinger (1996) showed that adults are better able to understand speech when presented by the same talker. Similarly, Kraljic et al. (2008) noted that adults adapt to speaker-specific idiosyncrasies. In particular, they showed that when presented with speech where the alveolar fricative "s" as in the word "see" was shifted to a more palatal place of articulation resembling "sh" as in "she", subjects were able to adapt and correctly identify the shifted "s" sounds — so long as they were provided with cues as to which variant of "s" was likely to occur. These situations are modeled by $\mathcal{M}_3$ and $\mathcal{M}_4$. $\mathcal{M}_3$ uses the additional cue y to help identify the talker, and hence, the correct distribution for the category over linguistic cue x. This way the indexical feature has an indirect impact on recognition even if there is no direct dependency between c and y. $\mathcal{M}_4$ attempts to adapt to the talker without the aid of the indexical cue. The model assumes such features exist, but are not observed and therefore cannot assist in identifying the talker. The prediction for $\mathcal{M}_4$ is that, like the subjects in the study by Kraljic et al. (2008), the model will perform more poorly and will incorrectly allow talker-specific variation to influence recognition of other talkers.

Inference

The models were implemented using WinBUGS (Spiegelhalter, Thomas, Best, & Lunn, 2003), which uses an automatic Gibbs sampling MCMC approach to estimate parameters and allows rapid prototyping and testing



$\mathcal{M}_1$	$\mathcal{M}_2$	$\mathcal{M}_3$	$\mathcal{M}_4$
$c \sim Bern(0.5)$	$c \sim Bern(0.5)$	$c \sim Bern(0.5)$	$c \sim Bern(0.5)$
		$t \sim Bern(0.5)$	$t \sim Bern(0.5)$
		$\lambda \sim \mathcal{N}(30, 5 \cdot 10^{-4})$	$\lambda \sim \mathcal{N}(30, 5 \cdot 10^{-4})$
		$\gamma \sim Gamma(0.2, 0.2)$	$\gamma \sim Gamma(0.2, 0.2)$
$\mu_c \sim \mathcal{N}(30, 5 \cdot 10^{-4})$	$\mu_c \sim \mathcal{N}(30, 5 \cdot 10^{-4})$	$\mu_{c,t} \vert \lambda_c, \gamma_c \sim \mathcal{N}(\lambda_c, \gamma_c)$	$\mu_{c,t} \vert \lambda_c, \gamma_c \sim \mathcal{N}(\lambda_c, \gamma_c)$
$\tau_c \sim Gamma(0.2, 0.2)$	$\tau_c \sim Gamma(0.2, 0.2)$	$\tau_{c,t} \sim Gamma(0.2, 0.2)$	$\tau_{c,t} \sim Gamma(0.2, 0.2)$
$x \vert c, \mu, \tau \sim \mathcal{N}(\mu_c, \tau_c)$	$x \vert c, \mu, \tau \sim \mathcal{N}(\mu_c, \tau_c)$	$x \vert c, t, \mu, \tau \sim \mathcal{N}(\mu_{c,t}, \tau_{c,t})$	$x \vert c, t, \mu, \tau \sim \mathcal{N}(\mu_{c,t}, \tau_{c,t})$
$y \vert c, \pi \sim Bern(\pi_c)$	$y \vert \pi \sim Bern(\pi)$	$y \vert t, \pi \sim Bern(\pi_t)$	$y \vert t, \pi \sim Bern(\pi_t)$

Figure 1: Four Possible Speech Perception Models: $\mathcal{M}_1$ treats all features as linguistic, $\mathcal{M}_2$ distinguishes the true and false linguistic features, $\mathcal{M}_3$ models individual talkers and treats some features as indexical, and $\mathcal{M}_4$ models talkers where the indexical features are absent or obscured. The variables are defined as follows: c is the speech sound category, t is the talker, x is a linguistic feature, y is an indexical feature, and the other variables are distributional parameters, defining talker and category specific distributions. C is the set of categories, T is the set of talkers, and S is the set of all speech sound tokens.

of Bayesian models.

We use an explicit initialization strategy, running the models in a generative mode with no observed variables and drawing category parameters for x at random from a $\mathcal{N}(50, 0.0025)$ for the mean and a $Gamma(2, 2)$ distribution for the precision. Using an initialization strategy such as this could speed convergence, since it tends to start the model out in a higher probability space. It also has the effect of reducing problems with numerical underflow error in WinBUGS. We were careful to pick the parameters randomly in such a way as to avoid biasing search in favor of any particular model or clustering, since we are primarily interested in the model properties, not the effects of initialization on convergence.

We find that even the more complex models converge in well under the 30,000 iterations we use. We average over the next 1000 iterations after convergence to measure the various parameters and statistics we report in subsequent sections. We take care in observing performance over these last 1000 iterations for any trends or abrupt changes. These mixture models have multiple symmetric optimal solutions, where “t” may be associated with cluster 1 and “d” with 2, or vice versa. If left to run long enough, the MCMC search strategy tends to switch between these different symmetric configurations every few thousand iterations. Averaging over instances of multiple such symmetric cases results in increased error in measurement. For instance, attempting to estimate the mean x value for phones in a cluster that toggles between “t” and “d” gets an average that is dissimilar to both configurations, and not only results in a measurement that is far from the gold standard but does not even accurately reflect the stationary distribution of the sampler.

Simulations

Data

We run the model on three synthetic data sets, illustrating the contrast between English word initial “t” and “d”. The primary difference between the two is in the voice onset time (VOT). We generate 100 sounds. Table 1 shows the model parameters used to generate each of the three data sets. For data set one we generate sounds as though there is only one speaker. For data set two we use two talkers, covarying the category with the talker so that instances of the first phone are produced by talker one and all instances of the second phone are produced by talker two. Finally, for data set three we split the 100 sounds evenly between the two talkers and the two categories, where talker and category are independent.

Simulation 1: The Developmental Situation

To simulate a situation similar to the psycholinguistic experiments of Houston and Jusczyk (2000), we present the models with two different data sets: data set one, where there is only one talker, and data set two, where there are two talkers, each producing just one of the two phones. In the behavioral experiment, it was observed that infants trained with word stimuli in a female voice were only able to reliably recognize words at test time when they were again presented in a female voice, and could not generalize to a male voice. Thus, the infants seem to confuse some non-linguistic feature of the sound, perhaps fundamental frequency, with the linguistic identity of the sounds. In this simulation, we shall

Table 1: Three Synthetic Data Sets

Talker	Parameter	Data Set		
		One	Two	Three
One	π_1	0.5	0.8	0.8
	π_2	0.5	0.8	0.8
	μ_1	15	15	0
	μ_2	35	-	35
	τ_1	15^{-2}	15^{-2}	15^{-2}
	τ_2	5^{-2}	-	5^{-2}
Two	π_1	-	0.2	0.2
	π_2	-	0.2	0.2
	μ_1	-	-	15
	μ_2	-	35	65
	τ_1	-	-	15^{-2}
	τ_2	-	5^{-2}	5^{-2}
Talkers Covary		-	Yes	No

Table 2: Categorization Accuracy

Model	Data Set		
	One	Two	Three
$\mathcal{M}_1$	0.77	**0.89**	0.52
$\mathcal{M}_2$	**0.81**	0.81	**0.76**

say that our indexical feature y corresponds to a thresholded fundamental frequency: sounds with a high fundamental frequency are more likely to be produced by the female talker, and lower fundamental frequency sounds by the male talker.

To simulate the developmental character of an infant's nascent linguistic capabilities, we perform a kind of structure discovery using Bayesian model selection between $\mathcal{M}_1$ and $\mathcal{M}_2$, where the infant is attempting to determine if the indexical feature y is relevant to the linguistic category ($\mathcal{M}_1$) or not ($\mathcal{M}_2$). We do this by introducing an additional latent variable corresponding to the model and define a uniform prior over the model. Then, we compute the probability of the model given the data, integrating out all other variables. To compare the two models, we simply compare the probabilities assigned to each model given the data. Typically, in such cases if the ratio $P(\mathcal{M}_1|D)/P(\mathcal{M}_2|D)$, called the Bayes factor, is greater than one, we say that model one is preferred, and otherwise model two is preferred.

In this case, whether we use data set one or two, virtually all the probability mass (approximately 100%) is placed on exactly one of the two models. $\mathcal{M}_1$ is overwhelmingly preferred when using data set two, the case where talker and phonetic category covary. On the other hand, data set one, the data set where both phonetic categories are produced by the same talker, results in an overwhelming preference for $\mathcal{M}_2$.

Table 2 presents accuracy results for the two models on the two data sets. Note that in general for these sorts of clustering algorithms there is an identifiability problem. That is, we cannot immediately say whether a particular category value $c = 1$ corresponds to the "t" or "d" sound. However, this poses less of a problem for this simple case with only two categories. For our purposes, it seems sufficient to assign the category that achieves highest accuracy.

We observe that while the model that mistakes the indexical for a linguistic feature ($\mathcal{M}_1$) performs very well for the artificially contrived covarying data, it performs worse on the data set that has only a single talker, and very nearly at chance for the data set with two different talkers that don't covary with the category. Figure 2 depicts the clusterings found by the two models on data set three (the data set with two talkers that don't covary with the phone). While $\mathcal{M}_2$ seems to do as well as can be hoped considering its inability to adapt to individual talkers, $\mathcal{M}_1$ very nearly fails to differentiate at all between "t" and "d". $\mathcal{M}_1$ attempts to cluster according to the indexical, collapsing the two categories together for each talker and clustering by talker instead of by category.

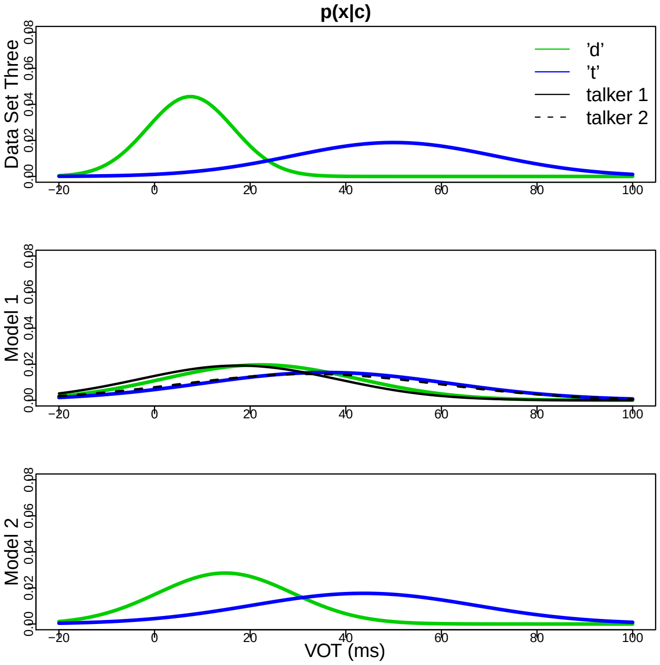


Figure 2: The conditional distribution over x given c for the true data set as compared to the two models $\mathcal{M}_1$ and $\mathcal{M}_2$. The clusters for the two talkers have been merged for ease of comparison. We also compare model 1's clusters against the distribution over x given the talker t.

Thus, the model selection approach predicts the psycholinguistic results very well. Training on sounds in one talker's voice, as in the covarying data set, results in the incorrect model being learned, which then fails to generalize to the same sound produced in the other talker's voice.

Simulation 2: Talker Adaptation

Adult talkers actually have the ability to adapt to individual talkers, learning to exploit talker specific variations

(Goldinger, 1996). To simulate this ability, we compare the performance of models $\mathcal{M}_3$ and $\mathcal{M}_4$. Model $\mathcal{M}_3$ corresponds to a case where the subject has learned that the indexical feature y can be used to identify the talker. On the other hand, $\mathcal{M}_4$ corresponds to the case where, although the subject is aware that the sounds may be produced by a different talker, the voice is disguised so that no cue is available for the identification of the talker. The contrast between these two models is similar to that demonstrated by Kraljic et al. (2008), where subjects were presented with ambiguous sounds that, in one condition, were accompanied by an additional cue indicating the ambiguity was result of talker dialect, and, in a second condition, were presented without this cue. This dialectical indicator, based on a phonological context, corresponds to our indexical feature y. Thus, condition one corresponds to $\mathcal{M}_3$ and condition two to $\mathcal{M}_4$. In the behavioral study, it was observed that subjects were much more prone to confusing the two different phonetic categories when the sounds were presented without the additional cue. Thus, we expect $\mathcal{M}_3$ to do much better.

Table 3 contains the categorization accuracy results for $\mathcal{M}_3$ and $\mathcal{M}_4$. Note that these models can theoretically identify the talker as well as the phonetic category, and we report accuracy for both. $\mathcal{M}_3$ does slightly better at clustering the phonetic categories, which is likely due to its much better ability to identify the talker. Note that without the indexical feature, $\mathcal{M}_4$ is at chance with regard to talker identification.

Table 3: Categorization Accuracy for Data Set Three

Model	Category	Talker
$\mathcal{M}_3$	**0.86**	**0.78**
$\mathcal{M}_4$	0.81	0.50

Figure 3 shows the clusters inferred by the two talker adapting models. The inferred Gaussian distributions for the two talkers are much more distinct for $\mathcal{M}_3$ than they are for $\mathcal{M}_4$ and more closely resemble the true distribution.

The inferred clusters, presented in Figure 3, are particularly interesting when compared against the findings of Kraljic et al. (2008), who observed that when the dialectical cue was absent, subjects adjusted their perceptual judgments for all talkers, not just the talker that produced the ambiguous variant. Model 3 makes use of the additional feature y for keeping the two talkers distinct, and therefore is less likely to let experience with the ambiguous talker influence its judgment for the other talker. Similarly, model 4 captures the situation where no additional cues are available. In this case, even if separate clusters are maintained for each talker, the two are functionally identical, falling somewhere in between the two true clusters. The mean is the mean of the two talker specific variants of the category, and, in the case of the "d" sound, the variance is much larger. Thus, the ambiguous talker influences recognition of the other talker when no additional cues are available, but not nearly as much when additional cues are discernible.

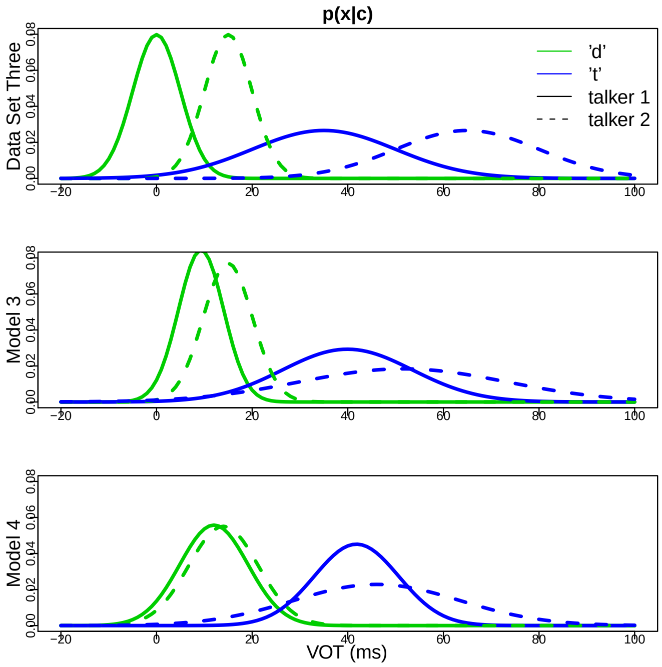


Figure 3: The conditional distribution over x given c for the true data set as compared to the two models $\mathcal{M}_3$ and $\mathcal{M}_4$.

As in the case of the developmental simulation, we see that the alternate performance of two models predicts the empirical results much better than would any one of the two.

Discussion and Conclusion

We have presented a computational model demonstrating a distributional account of certain covariance effects in infant and adult speech perception observed in the psycholinguistic literature. In particular, we found that by modeling the development of infant speech perception as a type of Bayesian model selection, we can account nicely for documented effects of covarying talker and phonetic category on infant confusions between categories (Houston & Jusczyk, 2000). We also found that by modeling talker identity, the same talker-specific features that confused the infant models could be exploited to improve performance, similar to demonstrations of talker adaptation in adult subjects (Kraljic et al., 2008). Also consistent with Kraljic et al. (2008), we found that when the talker adapting models were deprived of observed indexical information, talker specific speech habits influenced the category representations for all talkers not just the talker that produced the offending speech sounds.

While it would be difficult to account for all the phenomena with a single model of the statistical dependencies in the data, multiple models predict the empirical results fairly closely. This raises the question of how human subjects move between models, begging a model of the model selection process itself. Developmental shifts are readily handled in the Bayesian framework as a model selection problem, just the approach we took for explaining the infant behavior. Though it is beyond the scope of this paper, a similar selection process may account for a shift between the infant and adult stages, perhaps with several additional intermediate structures. We argue that modeling the developmental process as a shift between models rather than a gradual adjustment of a single model better matches the fact that there are distinct developmental stages. One set of models may correspond to a particular stage, where the underlying behavioral causes are made explicit by the dependency structure of the model.

Although the simulations we presented dealt primarily with covariance between talker and phonetic category, we expect that models based on similar principles could explain equally well other kinds of covariance phenomena, such as with speech affect and category (Singh, 2008). Note that the models we presented to explain infant phenomena had no explicit model of talker identity. Thus, the choice between the two models in the developmental case only constituted a feature selection task, where features that clearly covaried with the phonetic category were greatly preferred by the selection criterion. Thus, these simple models, in fact, generalize directly.

Similarly, while the talker adapting models do contain an explicit representation of talker identity, there is nothing that requires that the t variable refer to a talker. Similar variables could represent modes of talking, such as infant directed speech, or happy speech, or to dialectical variations or any number of other categorizable speech types. That is, the talker adapting models present a general adaptation strategy that could be employed with little or no modification.

We argue for the generality of the principles underlying our computational account while stressing that the full speech recognition problem, or even just that of phonetic category recognition, is a difficult one, and we have not attempted to model it in its entirety. In fact, we made several explicit simplifications. First, we assumed there are only two categories and two talkers. Second, we assumed that there are roughly equal numbers of tokens of each category, and that each talker produces about half of the sounds. Also, since we were primarily interested in how phonetic categories are learned, we assumed a simple Bernoulli distribution for the indexical feature, when, in fact, in many cases this feature too may very well be continuous. Furthermore, it was sufficient for our purposes to model a recognition problem along only one or two dimensions of the perceptual space.

These simplifications eased the implementation work but did not interfere with our ability to simulate the behavioral situations in which we were interested. They should not limit the generalizability of the results, and could be relaxed in a fairly straightforward manner if we wished to increase the realism. For instance, the first restriction could be relaxed by allowing the model to infer how many categories there are from the data using an infinite mixture model. We could also use a beta prior to infer relative talker and phonetic category frequency. A similar prior could be used to infer the distribution over the indexical features. Finally, multivariate Gaussians could be used for multiple correlated linguistic features (Vallabha, McClelland, Pons, Werker, & Amano, 2007). These are obvious extensions to consider for future work.

References

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.

Dorman, M. F., Studdert-Kennedy, M., & Raphael, L. J. (1977). Stop consonant recognition: Release bursts and formant transitions as functionally equivalent, context-dependent cues. *Perception & Psychophysics*, *22*, 109–122.

Espy-Wilson, C. Y. (1992). Acoustic measures for linguistic features distinguishing the semivowels /w j r l/ in american english. *Journal of the Acoustical Society of America*, *92*(1), 736–757.

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*, 1166–1183.

Houston, D. M., & Jusczyk, P. W. (2000). The role of talker-specific information in word segmentation by infants. *Journal of Experimental Psychology: Human Perception and Performance*, *26*(5), 1570–1582.

Kraljic, T., Brennan, S. E., & Samuel, A. G. (2008). Accommodating variation: Dialects, idiolects, and speech processing. *Cognition*, *107*, 54–81.

Lisker, L., & Abramson, A. A. (1964). A cross-language study of voicing in initial stops: acoustical measurements. *Word*, *20*, 384–422.

Magnuson, J. S., & Nusbaum, H. C. (2007). Acoustic differences, listener expectations, and the perceptual accommodation of talker variability. *Journal of Experimental Psychology: Human Perception and Performance*, *33*(2), 391–409.

Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, *82*(3), B101–B111.

McQueen, J. M., Norris, D., & Cutler, A. (2006). The dynamic nature of speech perception. *Language and Speech*, *49*, 101–112.

Peterson, G. E., & Barney, H. (1951). Control methods used in the study of vowels. *Journal of the Acoustical Society of America*, *24*(2), 175–184.

Singh, L. (2008). Influences of high and low variability on infant word recognition. *Cognition*, *106*, 833–870.

Spiegelhalter, D., Thomas, A., Best, N., & Lunn, D. (2003). *Winbugs: User manual, version 1.43*. Cambridge: Medical Research Council Biostatistics Unit.

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 Science*, *104*, 13273–13278.