Learning Phonetic Categories by Learning a Lexicon

Naomi H. Feldman (naomi_feldman@brown.edu)
Department of Cognitive and Linguistic Sciences, Brown University, Providence, RI 02912 USA

Thomas L. Griffiths (tom_griffiths@berkeley.edu)
Department of Psychology, University of California at Berkeley, Berkeley, CA 94720 USA

James L. Morgan (james_morgan@brown.edu)
Department of Cognitive and Linguistic Sciences, Brown University, Providence, RI 02912 USA

Abstract

Infants learn to segment words from fluent speech during the same period as they learn native language phonetic categories, yet accounts of phonetic category acquisition typically ignore information about the words in which speech sounds appear. We use a Bayesian model to illustrate how feedback from segmented words might constrain phonetic category learning, helping a learner disambiguate overlapping phonetic categories. Simulations show that information from an artificial lexicon can successfully disambiguate English vowel categories, leading to more robust category learning than distributional information alone.

Keywords: language acquisition; phonetic categories; Bayesian inference

Infants learning their native language need to extract several levels of structure, including the locations of phonetic categories in perceptual space and the identities of words they segment from fluent speech. It is often implicitly assumed that these steps occur sequentially, with infants first learning about the phonetic categories in their language and subsequently using those categories to help them map word tokens onto lexical items. However, infants begin to segment words from fluent speech as early as 6 months (Bortfeld, Morgan, Golinkoff, & Rathbun, 2005) and this skill continues to develop over the next several months (Jusczyk & Aslin, 1995; Jusczyk, Houston, & Newsome, 1999). Discrimination of non-native speech sound contrasts declines during the same time period, between 6 and 12 months (Werker & Tees, 1984). This suggests an alternative learning trajectory in which infants simultaneously learn to categorize both speech sounds and words, potentially allowing the two learning processes to interact.

In this paper we explore the hypothesis that the words infants segment from fluent speech can provide a useful source of information for phonetic category acquisition. We use a Bayesian approach to explore the nature of the phonetic category learning problem in an interactive system, where information from segmented words can feed back and constrain phonetic category learning. Our interactive model learns a rudimentary lexicon and a phoneme inventory[1] simultaneously, deciding whether acoustic representations of segmented tokens correspond to the same or different lexical items (e.g. bed vs. bad) and whether lexical items contain the same or different vowels (e.g. send vs. act). Simulations demonstrate that using information from segmented words to constrain phonetic category acquisition allows more robust category learning from fewer data points, due to the interactive learner's ability to use information about which words contain particular speech sounds to disambiguate overlapping categories.

The paper is organized as follows. We begin with an introduction to the mathematical framework for our model, then present toy simulations to demonstrate its qualitative properties. Next, simulations show that information from an artificial lexicon can disambiguate formant values associated with English vowel categories. The last section discusses potential implications for language acquisition, revisits the model's assumptions, and suggests directions for future research.

Bayesian Model of Phonetic Category Learning

Recent research on phonetic category acquisition has focused on the importance of distributional learning. Maye, Werker, and Gerken (2002) found that the specific frequency distribution (bimodal or unimodal) of speech sounds along a continuum could affect infants' discrimination of the continuum endpoints, with infants showing better discrimination of the endpoints when familiarized with the bimodal distribution. This work has inspired computational models that use a Mixture of Gaussians approach, assuming that phonetic categories are represented as Gaussian, or normal, distributions of speech sounds and that learners find the set of Gaussian categories that best represents the distribution of speech sounds they hear. Boer and Kuhl (2003) used the Expectation Maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977) to learn the locations of three such vowel categories from formant data. McMurray, Aslin, and Toscano (2009) introduced a gradient descent algorithm similar to EM to learn a stop consonant voicing contrast, and this algorithm has been extended to multiple dimensions for both consonant and vowel data (Toscano & McMurray, 2008; Vallabha, McClelland, Pons, Werker, & Amano, 2007).

Our model adopts the Mixture of Gaussians approach from these previous models but uses a non-parametric Bayesian framework that allows extension of the model to the word level, making it possible to investigate the learning outcome when multiple levels of structure interact. As in previous models, speech sounds in our model are represented using

[1] We make the simplifying assumption that phonemes are equivalent to phonetic categories, and use the terms interchangeably.

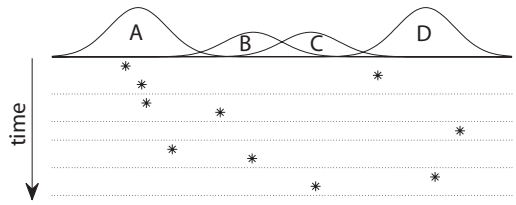

Figure 1: A fragment of a corpus presented to the model. Asterisks represent speech sounds, and lines represent word boundaries. The model does not know which categories generated the speech sounds, and needs to recover categories A, B, C, and D from the data.

phonetic dimensions such as steady-state formant values or voice onset time. Words are sequences of these phonetic values, where each phoneme corresponds to a single discrete set (e.g. first and second formant) of phonetic values. A sample fragment of a toy corpus is shown in Figure 1. The phoneme inventory has four categories, labeled A, B, C, and D; five words are shown, representing lexical items ADA, AB, D, AB, and DC, respectively. Learning involves using the speech sounds and, in the case of an interactive learner, information about which other sounds appear with them in words, to recover the phonetic categories that generated the corpus.

Simulations compare two models that differ in the hypothesis space they assign to the learner. In the distributional model, the learner's hypothesis space contains phoneme inventories, where phonemes correspond to Gaussian distributions of speech sounds in phonetic space. In the lexical-distributional model, the learner considers these same phoneme inventories, but considers them only in conjunction with lexicons that contain lexical items composed of sequences of phonemes. This allows the lexical-distributional learner to use not only phonetic information, but also information about the words that contain those sounds, in recovering a set of phonetic categories.

Distributional Model

In the distributional model, a learner is responsible for recovering a set of phonetic categories, which we refer to as a phoneme inventory C, from a corpus of speech sounds. The model ignores all information about words and word boundaries, and learns only from the distribution of speech sounds in phonetic space. Speech sounds are assumed to be produced by selecting a phonetic category c from the phoneme inventory and then sampling a phonetic value from the Gaussian associated with that category. Categories differ in their means μ_c, covariance matrices Σ_c, and frequencies of occurrence.

Following previous work in morphology (Goldwater, Griffiths, & Johnson, 2006), word segmentation (Goldwater, Griffiths, & Johnson, in press), and grammar learning (Johnson, Griffiths, & Goldwater, 2007), learners' prior beliefs about the phoneme inventory are encoded using a nonparametric Bayesian model called the Dirichlet process (Ferguson, 1973), $C \sim DP(\alpha, G_C)$. This distribution encodes biases over the number of categories in the phoneme inventory, as well as over phonetic parameters for those categories. Prior beliefs about the number of phonetic categories allow the learner to consider a potentially infinite number of categories, but produce a bias toward fewer categories, with the strength of the bias controlled by the parameter α.[2] This replaces the winner-take-all bias in category assignments that has been used in previous models (McMurray et al., 2009; Vallabha et al., 2007) and allows explicit inference of the number of categories needed to represent the data.

The prior distribution over phonetic parameters is defined by G_C, which in this model is a distribution over Gaussian phonetic categories that includes an Inverse-Wishart prior over category variances, $\Sigma_c \sim IW(\nu_0, \Sigma_0)$, and a Gaussian prior over category means, $\mu_c|\Sigma_c \sim N(\mu_0, \frac{\Sigma_c}{\nu_0})$. The parameters of these distributions can be thought of as pseudodata, where μ_0, Σ_0, and ν_0 encode the mean, covariance, and number of speech sounds that the learner imagines having already assigned to any new category. This prior distribution over phonetic parameters is not central to the theoretical model, but rather is included for ease of computation; the number of speech sounds in the pseudodata is made as small as possible[3] so that the prior biases are overshadowed by real data.

Presented with a sequence of acoustic values, the learner needs to recover the set of Gaussian categories that generated those acoustic values. Gibbs sampling (Geman & Geman, 1984), a form of Markov chain Monte Carlo, is used to recover examples of phoneme inventories that an ideal learner believes are likely to have generated the corpus. Speech sounds are initially given random category assignments, and in each sweep through the corpus, each speech sound in turn is given a new category assignment based on all the other current assignments. The probability of assignment to category c is given by Bayes' rule,

$$p(c|w_{ij}) \propto p(w_{ij}|c)p(c) \tag{1}$$

where w_{ij} denotes the phonetic parameters of the speech sound in position j of word i. The prior $p(c)$ is given by the Dirichlet process and is

$$p(c) = \begin{cases} \frac{n_c}{\sum_c n_c + \alpha} & \text{for existing categories} \\ \frac{\alpha}{\sum_c n_c + \alpha} & \text{for a new category} \end{cases} \tag{2}$$

making it proportional to the number of speech sounds n_c already assigned to that category, with some probability α of assignment to a new category. The likelihood $p(w_{ij}|c)$ is obtained by integrating over all possible means and covariance matrices for category c, $\int\int p(w_{ij}|\mu_c, \Sigma_c)p(\mu_c|\Sigma_c)p(\Sigma_c)d\mu_c d\Sigma_c$, where the probability distributions $p(\mu_c|\Sigma_c)$ and $p(\Sigma_c)$ are modified to take into account the speech sounds already assigned to that category.

[2] This bias is needed to induce any grouping at all; the maximum likelihood solution assigns each speech sound to its own category.

[3] To form a proper distribution, ν_0 needs to be greater than $d-1$, where d is the number of phonetic dimensions.

This likelihood function has the form of a multivariate t-distribution and is discussed in more detail in Gelman, Carlin, Stern, and Rubin (1995). Using this procedure, category assignments converge to the posterior distribution on phoneme inventories, revealing an ideal learner's beliefs about which categories generated the corpus.

Lexical-Distributional Model

This non-parametric Bayesian framework has the advantage that it is straightforward to extend to hierarchical structures (Teh, Jordan, Beal, & Blei, 2006), allowing us to explore the influence of words on phonetic category acquisition. In the lexical-distributional model, the learner recovers not only the same phoneme inventory C as in the distributional model, but also a lexicon L with lexical items composed of sequences of phonemes. This creates an extra step in the generative process: instead of assuming that the phoneme inventory generates a corpus directly, as in the distributional model, this model assumes that the phoneme inventory generates the lexicon and that the lexicon generates the corpus. The corpus is generated by selecting a lexical item to produce and then sampling an acoustic value from each of the phonetic categories contained in that lexical item.

The prior probability distribution over possible lexicons is a second Dirichlet process, $L \sim DP(\beta, G_L)$ where G_L defines a prior distribution over lexical items. This prior favors shorter lexical items, assuming word lengths to be generated from a geometric distribution, and assumes that a category for each phoneme slot has been sampled from the phoneme inventory C. Thus, the prior probability distribution over words is defined according to the phoneme inventory, and the learner needs to optimize the phoneme inventory so that it generates the lexicon. Parallel to the bias toward fewer phonetic categories, the model encodes a bias toward fewer lexical items but allows a potentially infinite number of lexical items.

Presented with a corpus consisting of isolated word tokens, each of which consists of a sequence of acoustic values, the language learner needs to recover the lexicon and phoneme inventory of the language that generated the corpus. Learning is again performed through Gibbs sampling. Each iteration now includes two sweeps: one through the corpus, assigning each word to the lexical item that generated it, and one through the lexicon, assigning each position of each lexical item to its corresponding phoneme from the phoneme inventory. In the first sweep we use Bayes' rule to calculate the probability that word w_i corresponds to lexical item k,

$$p(k|w_i) \propto p(w_i|k)p(k) \tag{3}$$

Parallel to Equation 2, the prior is

$$p(k) = \begin{cases} \frac{n_k}{\sum_k n_k + \beta} & \text{for existing categories} \\ \frac{\beta}{\sum_k n_k + \beta} & \text{for a new category} \end{cases} \tag{4}$$

where n_k is the number of word tokens already assigned to lexical item k. A word is therefore assigned to a lexical item with a probability proportional to the number of times that lexical item has already been seen, with some probability β reserved for the possibility of seeing a new lexical item. The likelihood is a product of the likelihoods of each speech sound having been generated from its respective category,

$$p(w_i|k) = \prod_j p(w_{ij}|c_{kj}) \tag{5}$$

where j indexes a particular position in the word and c_{kj} is the phonetic category that corresponds to position j of lexical item k. Any lexical item with a different length from the word w_i is given a likelihood of zero, and samples from the prior distribution on lexical items are used to estimate the likelihood of a new lexical item (Neal, 1998).

The second sweep uses Bayes' rule

$$p(c|w_{\{k\}j}) \propto p(w_{\{k\}j}|c)p(c) \tag{6}$$

to assign a phonetic category to position j of lexical item k, where $w_{\{k\}j}$ is the set of phonetic values at position j in all of the words in the corpus that have been assigned to lexical item k. The prior $p(c)$ is the same prior over category assignments as was used in the distributional model, and is given by Equation 2. The likelihood $p(w_{\{k\}j}|c)$ is again computed by integrating over all possible means and covariance matrices, $\int\int \prod_{w_i \in k} p(w_{ij}|\mu_c, \Sigma_c)p(\mu_c|\Sigma_c)p(\Sigma_c)d\mu_c d\Sigma_c$, this time taking into account phonetic values from all the words assigned to lexical item k. The sampling procedure converges on samples from the joint posterior distribution on lexicons and phoneme inventories, allowing learners to recover both levels of structure simultaneously.

Qualitative Behavior of an Interactive Learner

In this section, toy simulations demonstrate how a lexicon can provide disambiguating information about overlapping categories that would be interpreted as a single category by a purely distributional learner. We show that it is not the simple presence of a lexicon, but rather specific disambiguating information within the lexicon, that increases the robustness of category learning in the lexical-distributional learner.

Corpora were constructed for these simulations using four categories labeled A, B, C, and D, whose means are located at -5, -1, 1, and 5 along an arbitrary phonetic dimension (Figure 2 (a)). All four categories have a variance of 1. Because the means of categories B and C are so close together, being separated by only two standard deviations, the overall distribution of tokens in these two categories is unimodal.

To test the distributional learner, 1200 acoustic values were sampled from these categories, with 400 acoustic values sampled from each of Categories A and D and 200 acoustic values sampled from each of Categories B and C. Results indicate that these distributional data are not strong enough to disambiguate categories B and C, leading the learner to interpret them as a single category (Figure 2 (b)).[4] While this may

[4]Simulations in this section used parameters $\alpha = \beta = 1$, $\mu_0 = 0$, $\Sigma_0 = 1$, and $\nu_0 = 0.001$; each simulation was run for 500 iterations.

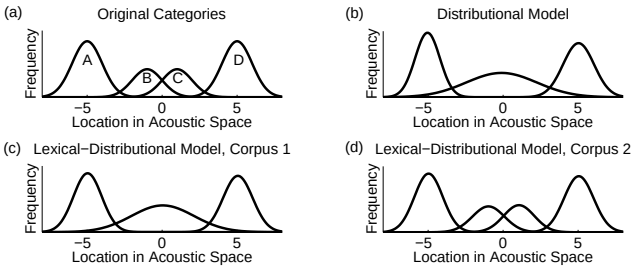

Figure 2: Toy data with two overlapping categories as (a) generated, (b) learned by the distributional model, (c) learned by the lexical-distributional model from a minimal pair corpus, and (d) learned by the lexical-distributional model from a corpus without minimal pairs.

be due in part to the distributional learner's prior bias toward fewer categories, simulations in the next section will show that the gradient descent learner from Vallabha et al. (2007), which has no such explicit bias, shows similar behavior.

Two toy corpora were constructed for the lexical-distributional model from the 1200 phonetic values sampled above. The corpora differed from each other only in the distribution of these values across lexical items. The lexicon of the first corpus contained no disambiguating information about speech sounds B and C. It was generated from six lexical items, with identities AB, AC, DB, DC, ADA, and D. Each lexical item was repeated 100 times in the corpus for a total of 600 word tokens. In this corpus, Categories B and C appeared only in minimal pair contexts, since both AB and AC, as well as both DB and DC, were words. As shown in Figure 2 (c), the lexical-distributional learner merged categories B and C when trained on this corpus. Merging the two categories allowed the learner to condense AB and AC into a single lexical item, and the same happened for DB and DC. Because the distribution of these speech sounds in lexical items was identical, lexical information could not help disambiguate the categories.

The second corpus contained disambiguating information about categories B and C. This corpus was identical to the first except that the acoustic values representing the phonemes B and C of words AC and DB were swapped, converting these words into AB and DC, respectively. Thus, the second corpus contained only four lexical items, AB, DC, ADA, and D, and there were now 200 tokens of words AB and DC. Categories B and C did not appear in minimal pair contexts, as there was a word AB but no word AC, and there was a word DC but no word DB. The lexical-distributional learner was able to use the information contained in the lexicon in the second corpus to successfully disambiguate categories B and C (Figure 2 (d)). This occurred because the learner could categorize words AB and DC as two different lexical items simply by recognizing the difference between categories A and D, and could use those lexical classifications to notice small phonetic differences between the second phonemes in these lexical items.

In this model it is non-minimal pairs, rather than minimal pairs, that help the lexical-distributional learner disambiguate phonetic categories. While minimal pairs may be useful when a learner knows that two similar sounding tokens have different referents, they pose a problem in this model because the learner hypothesizes that similar sounding tokens represent the same word. Thiessen (2007) has made a similar observation with 15-month-olds in a word learning task, showing that infants may fail to notice a difference between similar-sounding object labels, but are better at discriminating these words when familiarized with non-minimal pairs that contain the same sounds.

Learning English Vowels

The prototypical examples of overlapping categories in natural language are vowel categories, such as the English vowel categories from Hillenbrand, Getty, Clark, and Wheeler (1995) shown in Figure 4 (a).[5] We therefore use English vowel categories to test the lexical-distributional learner's ability to disambiguate overlapping categories that are based on actual phonetic category parameters.

Two corpora were constructed using phonetic categories based on the Hillenbrand et al. (1995) vowel formant data. Categories in the first corpus were based on vowels spoken by men, and had only moderate overlap (Figure 3 (a)); categories in the second corpus were based on vowels spoken by men, women, and children, and had a much higher degree of overlap (Figure 4 (a)). In each case, means and covariance matrices for the twelve phonetic categories were computed from corresponding vowel tokens. Using the generative model, a hypothetical set of lexical items consisting only of vowels was generated for each corpus, and 5,000 word tokens were generated based on this lexicon from the appropriate set of Gaussian category parameters.

These corpora were given as training data to three models: the lexical-distributional model, the distributional model, and the multidimensional gradient descent algorithm used by Vallabha et al. (2007).[6] Results for the corpus based on men's productions are shown in Figure 3, and results from the corpus based on all speakers' productions are shown in Figure 4. In each case, the lexical-distributional learner recovered the correct set of vowel categories and successfully disambiguated neighboring categories. In contrast, the models lacking a lexicon mistakenly merged several pairs of neigh-

[5]These vowel data were obtained through download from http://homepages.wmich.edu/~hillenbr/.

[6]Parameters for the Bayesian models were $\alpha = \beta = 1$, $\mu_0 = \begin{bmatrix} 500 \\ 1500 \end{bmatrix}$, $\Sigma_0 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, and $\nu_0 = 1.001$, and each simulation was run for 600 iterations. No attempt was made to optimize these parameters, and they were actually different from the parameters used to generate the data, as $\alpha = \beta = 10$ was used to help produce a corpus that contained all twelve vowel categories. Using the generating parameters during inference did not qualitatively affect the results. Parameters for the gradient descent algorithm were identical to those used by Vallabha et al. (2007); optimizing the learning rate parameter produced little qualitative change in the learning outcome.

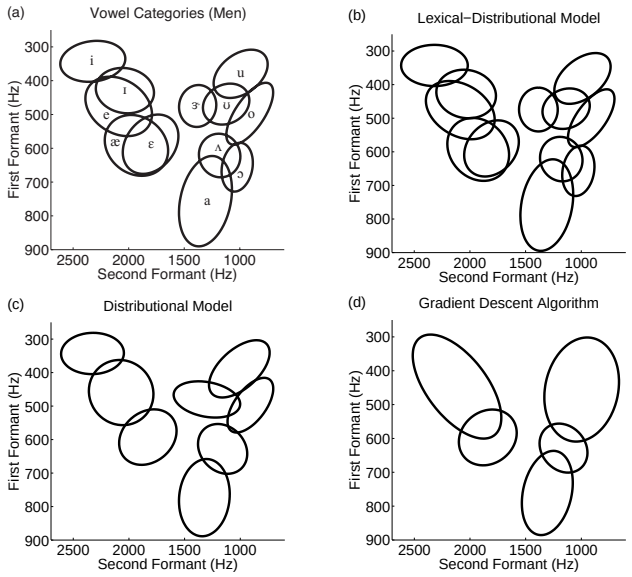

Figure 3: Ellipses delimit the area corresponding to 90% of vowel tokens for Gaussian categories (a) computed from men's vowel productions from Hillenbrand et al. (1995) and learned by the (b) lexical-distributional model, (c) distributional model, and (d) gradient descent algorithm.

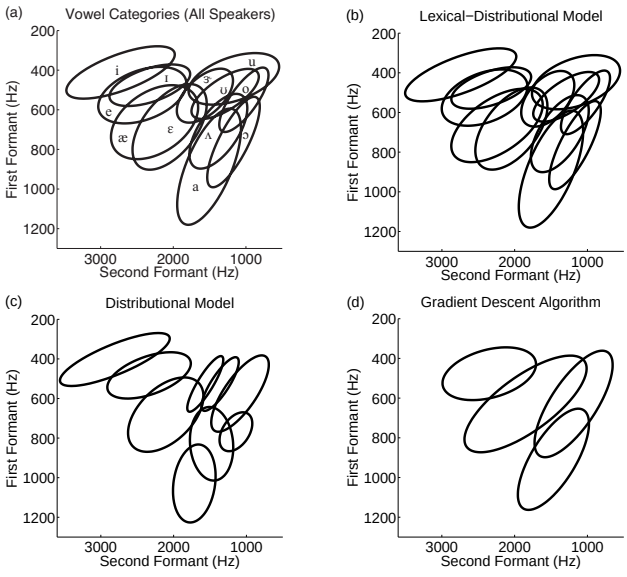

Figure 4: Ellipses delimit the area corresponding to 90% of vowel tokens for Gaussian categories (a) computed from all speakers' vowel productions from Hillenbrand et al. (1995) and learned by the (b) lexical-distributional model, (c) distributional model, and (d) gradient descent algorithm.

boring vowel categories. Positing the presence of a lexicon therefore showed evidence of helping the ideal learner disambiguate overlapping vowel categories, even though the phonological forms contained in the lexicon were not given explicitly to the learner.

Pairwise accuracy and completeness measures were computed for each learner as a quantitative measure of model performance (Table 1). For these measures, pairs of vowel tokens that were correctly placed into the same category were counted as a *hit*; pairs of tokens that were incorrectly assigned to different categories when they should have been in the same category were counted as a *miss*; and pairs of tokens that were incorrectly assigned to the same category when they should have been in different categories were counted as a *false alarm*. The accuracy score was computed as $\frac{\text{hits}}{\text{hits+false alarms}}$ and the completeness score as $\frac{\text{hits}}{\text{hits+misses}}$. Both measures were high for the lexical-distributional learner, but accuracy scores were substantially lower for the purely distributional learners, reflecting the fact that these models mistakenly merged several overlapping categories.

Results suggest that as predicted, a model that uses the input to learn word categories in addition to phonetic categories produces better phonetic category learning results than a model that only learns phonetic categories. Note that the distributional learners are likely to show better performance if they are given dimensions beyond just the first two formants (Vallabha et al., 2007) or if they are given more data points during learning. These two solutions actually work against each other: as dimensions are added, more data are necessary to maintain the same learning outcome. Nevertheless, we do not wish to suggest that a purely distributional learner cannot acquire phonetic categories. The simulations presented here are instead meant to demonstrate that in a language where phonetic categories have substantial overlap, an interactive system, where learners can use information from words that contain particular speech sounds, can increase the robustness of phonetic category learning.

Discussion

This paper has presented a model of phonetic category acquisition that allows interaction between speech sound and word categorization. The model was not given a lexicon a priori, but was allowed to begin learning a lexicon from the data at the same time that it was learning to categorize individual speech sounds, allowing it to take into account the distribution of speech sounds in words. This lexical-distributional learner outperformed a purely distributional learner on a corpus whose categories were based on English vowel categories, showing better disambiguation of overlapping categories from the same number of data points.

Infants learn to segment words from fluent speech around the same time that they begin to show signs of acquiring native language phonetic categories, and they are able to map these segmented words onto tokens heard in isolation (Jusczyk & Aslin, 1995), suggesting that they are performing some sort of rudimentary categorization on the words they hear. Infants may therefore have access to information from words that can help them disambiguate overlapping categories. If information from words can feed back to constrain phonetic category learning, the large degree of overlap be-

		Lexical-Distrib.	Distrib.	Gradient Descent
(a)	Accuracy	0.97	0.63	0.56
	Completeness	0.98	0.93	0.94
(b)	Accuracy	0.99	0.54	0.40
	Completeness	0.99	0.85	0.95

Table 1: Accuracy and completeness scores for learning vowel categories based on productions by (a) men and (b) all speakers. For the Bayesian learners, these were computed at the annealed solutions; for the gradient descent learner, they were based on maximum likelihood category assignments.

tween phonetic categories may not be such a challenge as is often supposed.

In generalizing these results to more realistic learning situations, however, it is important to take note of two simplifying assumptions that were present in our model. The first key assumption is that speech sounds in phonetic categories follow the same Gaussian distribution regardless of phonetic or lexical context. In actual speech data, acoustic characteristics of sounds change in a context-dependent manner due to coarticulation with neighboring sounds (e.g. Hillenbrand, Clark, & Nearey, 2001). A lexical-distributional learner hearing reliable differences between sounds in different words might erroneously assign coarticulatory variants of the same phoneme to different categories, having no other mechanism to deal with context-dependent variability. Such variability may need to be represented explicitly if an interactive learner is to categorize coarticulatory variants together.

A second assumption concerns the lexicon used in the vowel simulations, which was generated from our model. Generating a lexicon from the model ensured that the learner's expectations about the lexicon matched the structure of the lexicon being learned, and allowed us to examine the influence of lexical information in the best case scenario. However, several aspects of the lexicon, such as the assumption that phonemes in lexical items are selected independently of their neighbors, are unrealistic for natural language. In future work we hope to extend the present results using a lexicon based on child-directed speech.

Infants learn multiple levels of linguistic structure, and it is often implicitly assumed that these levels of structure are acquired sequentially. This paper has instead investigated the optimal learning outcome in an interactive system using a non-parametric Bayesian framework that permits simultaneous learning at multiple levels. Our results demonstrate that information from words can lead to more robust learning of phonetic categories, providing one example of how such interaction between domains might help make the learning problem more tractable.

Acknowledgments. This research was supported by NSF grant BCS-0631518, AFOSR grant FA9550-07-1-0351, and NIH grant HD32005. We thank Joseph Williams for help in working out the model and Sheila Blumstein, Adam Darlow, Sharon Goldwater, Mark Johnson, and members of the computational modeling reading group for helpful comments and discussion.

References

Boer, B. de, & Kuhl, P. K. (2003). Investigating the role of infant-directed speech with a computer model. *Acoustics Research Letters Online*, *4*(4), 129-134.

Bortfeld, H., Morgan, J. L., Golinkoff, R. M., & Rathbun, K. (2005). Mommy and me: Familiar names help launch babies into speech-stream segmentation. *Psychological Science*, *16*(4), 298-304.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, B*, *39*, 1-38.

Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, *1*(2), 209-230.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian data analysis*. New York: Chapman and Hall.

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE-PAMI*, *6*, 721-741.

Goldwater, S., Griffiths, T. L., & Johnson, M. (2006). Interpolating between types and tokens by estimating power-law generators. *Advances in Neural Information Processing Systems 18*.

Goldwater, S., Griffiths, T. L., & Johnson, M. (in press). A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition*.

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.

Hillenbrand, J. L., Clark, M. J., & Nearey, T. M. (2001). Effects of consonant environment on vowel formant patterns. *Journal of the Acoustical Society of America*, *109*(2), 748-763.

Johnson, M., Griffiths, T. L., & Goldwater, S. (2007). Adaptor grammars: a framework for specifying compositional nonparametric Bayesian models. *Advances in Neural Information Processing Systems 19*.

Jusczyk, P. W., & Aslin, R. N. (1995). Infants' detection of the sound patterns of words in fluent speech. *Cognitive Psychology*, *29*, 1-23.

Jusczyk, P. W., Houston, D. M., & Newsome, M. (1999). The beginnings of word segmentation in English-learning infants. *Cognitive Psychology*, *39*, 159-207.

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

McMurray, B., Aslin, R. N., & Toscano, J. C. (2009). Statistical learning of phonetic categories: Computational insights and limitations. *Developmental Science*, *12*(3), 369-378.

Neal, R. M. (1998). Markov chain sampling methods for Dirichlet process mixture models. *Technical Report No. 9815, Department of Statistics, University of Toronto*.

Teh, Y. W., Jordan, M. I., Beal, M. J., & Blei, D. M. (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, *101*, 1566-1581.

Thiessen, E. D. (2007). The effect of distributional information on children's use of phonemic contrasts. *Journal of Memory and Language*, *56*(1), 16-34.

Toscano, J. C., & McMurray, B. (2008). Using the distributional statistics of speech sounds for weighting and integrating acoustic cues. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (p. 433-438). Austin, TX: Cognitive Science Society.

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*, *104*, 13273-13278.

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