Learning distributions as they come: Particle filter models for online distributional learning of phonetic categories

Dave F. Kleinschmidt (davidfk@princeton.edu)
Princeton Neuroscience Institute
Princeton, NJ 08544 USA

Abstract

Human infants have the remarkable ability to learn any human language. One proposed mechanism for this ability is distributional learning, where learners infer the underlying cluster structure from unlabeled input. Computational models of distributional learning have historically been principled but psychologically-implausible computational-level models, or ad hoc but psychologically plausible algorithmic-level models. Approximate rational models like particle filters can potentially bridge this divide, and allow principled, but psychologically plausible models of distributional learning to be specified and evaluated. As a proof of concept, I evaluate one such particle filter model, applied to learning English voicing categories from distributions of voice-onset times (VOTs). I find that this model learns well, but behaves somewhat differently from the standard, unconstrained Gibbs sampler implementation of the underlying rational model.

Keywords: Computational modeling; Rational models; Particle filters; Language learning; Distributional learning; Speech perception

Any normally developing human infant can learn any human language, and they can do so without requiring much explicit instruction or guidance. How is this possible? At a basic level, how can an infant figure out that their language has, for instance, two different phonetic categories distinguished by voicing (e.g., "beach" vs. "peach"), and not three, or just one (Lisker & Abramson, 1964)? One clue comes from the fact that sounds from the same category tend to sound alike, more so than sounds from different categories. Infants (and adults) are sensitive to this cluster structure, and even in the absence of explicit cues or instructions learn to distinguish between two sounds better when they occur in two different clusters than when they occur in a single, unimodal cluster (Maye, Werker, & Gerken, 2002).

There are a number of different models for this "distributional learning", which fall into two broad families. On the one hand, there are *computational-level* models (in the sense of Marr, 1982), which focus on the nature of the problem to be solved, the information that is available from the world, and the best performance that is possible given the combination of those two factors (e.g., Feldman, Griffiths, Goldwater, & Morgan, 2013). On the other hand, there are cognitive, *algorithmic-level* models, which focus on psychologically plausible representations and processes (e.g., McMurray, Aslin, & Toscano, 2009; Vallabha, McClelland, Pons, Werker, & Amano, 2007). Both of these approaches have provided insight into the process of distributional learning. Computational-level models set the boundaries on what is possible *in principle*. A notable example is the work of Feldman et al. (2013) which shows, somewhat counterintuitively, that distributional learning of phonetic categories can in general be *enhanced* by simultaneous distributional learning of words. However, the implementations of these models are often profoundly psychologically implausible, and may assume that learners have simultaneous access to the entire batch of data to learn from, can make multiple passes over that data, or maintain unlimited amounts of uncertainty.

Algorithmic-level models, on the other hand, serve as existence proofs that distributional learning is possible, given particular *representational* assumptions. These models often make ad hoc assumptions in order to better fit behavioral data, as in McMurray et al. (2009) who conclude that "winner-take-all" competitive dynamics are necessary for distributional learning. These models also generally lack the in-principle guarantees of computational-level models, and thus it is unclear whether any particular model's failure or success reflects fundamental constraints or representational assumptions (though computational-level models are not immune to this problem; Goldwater, Griffiths, & Johnson, 2009).

Each level of analysis offers insight, but bridging between these levels is critical for a comprehensive understanding of human cognition (Marr, 1982). One way to approach such a bridge is the family of models known as rational approximations of rational models (Sanborn, Griffiths, & Navarro, 2010). These models provide a principled link between computational-level concerns about the structure of the world and the nature of the tasks that the mind must solve, and algorithmic-level concerns about psychologically plausible representations and processes. In this paper, I explore a psychologically plausible, approximately rational model for phonetic distributional learning: particle filters for Bayesian nonparametric clustering. As a case study, I apply this model to English stop voicing (e.g., /b/ vs. /p/), and investigate how the performance of this model is affected by the constraints of online processing (one data point at a time) and limited representational resources, relative to an unconstrained distributional learning algorithm.

To start, I review the Bayesian approach to distributional learning as a problem of statistical inference under uncertainty. Next, I briefly summarize two algorithms for doing this inference: a batch Gibbs sampler algorithm, and an online particle filter algorithm. I then analyze the performance of these two models on the same (simulated) phonetic cue distributions, focusing on how well they can recover the true underlying structure, before discussing the implications of these results.

Bayesian models of distributional learning

Distributional learning models all attempt to solve the same problem: given some data points $x_{1:N}$, we wish to know how many clusters generated them (and what those clusters look like). We assume that these data were generated from some number of clusters K, where K could be (in principle) anywhere between 1 and N (the number of data points). A complete clustering of these data points has two parts: the cluster that each data point is assigned to (denoted $c_{1:N}$), and the properties of the clusters themselves (which in the example below are the mean μ_k and variance σ_k^2 of a normal distribution, but in general are the parameters of some probability distribution).

The first difficulty is that there's no perfect, un-ambiguous solution to this problem (in general), since clusters are often overlapping and of different sizes and frequencies. For this reason, Bayesian nonparametric models frame this as a problem of *statistical inference under uncertainty*, and aim to find the posterior probability for each possible clustering, given some particular data, $p(c_{1:N}, \mu_{1:K}, \sigma^2_{1:K} | x_{1:N})$ (see Gershman & Blei, 2012, for an excellent introduction). In many cases, it's easier to do this in two stages: first, compute the distribution of cluster assignments $p(c_{1:N}|x_{1:N})$ (marginalizing over possible clusters *parameters*), and then compute the distribution over cluster properties, conditional on the cluster assignments $p(\mu_{1:K}, \sigma^2_{1:K} | c_{1:N}, x_{1:N})$.[1] Here the focus is on the first part: computing the distribution of cluster assignments given the data.

In principle, computing the posterior distribution $p(c_{1:N}|x_{1:N})$ is a simple matter of applying Bayes Rule:

$$p(c_{1:N}|x_{1:N}) \propto p(x_{1:N}|c_{1:N})p(c_{1:N})$$

The first term is the likelihood, or the probability that the observed data is generated by a particular hypothetical clustering, which will depend on how similar the data points are in each cluster, and on the prior distribution over cluster parameters. The second term is the prior, or how likely such a clustering is before any data is observed. The prior is where inductive biases for simpler clustering can be introduced.

The particular prior used here is the Dirichlet process. This prior has a rich-get-richer structure, which is most easily understood by considering the conditional prior $p(c_n|c_{1:n-1})$ over cluster assignments for the nth data point, given assignments 1 to $n-1$. The prior probability of assigning to a cluster k is

$$p(c_n = k | c_{1:n-1}) \propto \begin{cases} N_k & \text{existing cluster} \\ \alpha & \text{new cluster} \end{cases}$$

where N_k is the the number of other data points assigned to cluster k. The α parameter is called the concentration parameter and controls how strong the simplicity bias is. If α is very low, it's highly unlikely (a priori) that a new cluster will be created, especially when there are many data points. The overall prior probability of any complete clustering $c_{1:N}$ can be computed from these conditional probabilities under the assumption that the data is *exchangeable* (that order doesn't matter, or that the cluster structure is stable over time), and is proportional to $\prod_{2=1}^{N} p(c_i|c_{1:i-1}) \propto \prod_{k=1}^{K} \alpha N_k!$ (arbitrarily defining $c_1 = 1$).

The second difficulty is that there are an enormous number of possible combinations of cluster assignments. Even if we know that there are only $K = 2$ clusters, there are still 2^N possible ways to assign N points to two clusters, each of which has some probability associated with it. Likewise for every K from 1 to N. Luckily, most of these values for $c_{1:N}$ have probability that is so small it's essentially zero, and thus the whole distribution $p(c_{1:N}|x_{1:N})$ can be approximated by a reasonably small number of *samples*, or hypothetical values of $c_{1:N}$.

Batch algorithm: Gibbs sampler

One standard method of doing approximate inference by sampling is a Gibbs sampler, a form of Markov Chain Monte Carlo (MCMC) techniques. These are named because they work by sampling (the "Monte Carlo" part) a new value for the quantity of interest given only the data and the previously sampled value (the "Markov Chain" part).

For a Dirichlet Process mixture model, this algorithm works by sweeping through the data, one data point at a time, re-sampling the cluster assignment for that data point conditioned on the other data points. If c_{-i} are the cluster assignments for every point but x_i, then the Gibbs sampler will assign x_i to cluster k with probability $p(c_i = k|x_{1:N}, c_{-i}) \propto p(x_i|c_i = k, x_{-i}, c_{-i})p(c_i = k|c_{-i})$—that is, proportional to the likelihood of x_i given the other data points in cluster k times the prior probability of k under the Chinese Restaurant Process prior. The likelihood for a new category is based on the prior for the category parameters (e.g., mean and variance). Once new assignments have been sampled for every x_i, the new values of c_i are one sample from the posterior. Multiple samples are drawn by repeatedly sweeping through the data in this way, recording the sampled assignments for each sweep.

Online algorithm: Particle filter

In contrast to MCMC algorithms, sequential Monte Carlo (SMC) algorithms do not require that all data be available simultaneously. Rather then generating samples one at a time based on the entire dataset, they maintain a population of hypotheses (or particles), that are each updated in parallel as the data comes in. After $n-1$ observations, particle j has an associated weight $w^{(j)}_{n-1}$ and clustering $c^{(j)}_{1:n-1}$. There are many different strategies for updating particle j based on the the next observation x_n. Here, we follow the approach of Chen and Liu (2000), as described in, Fearnhead (2004). First, as in the Gibbs sampler, an assignment is drawn from

[1] One reason for this is that often when using a conjugate prior for the cluster parameters, there's no need to actually know the exact properties of cluster to evaluate how good it is; all that's needed is the data points that are assigned to that cluster (and some measure of how similar they are to each other).

$p(c_n = k|c^{(j)}_{1:n-1}, x_{1:n})$. Next, the weight is updated to be

$$w^{(j)}_n = w^{(j)}_{n-1} \times \frac{\sum_k p((c^{(j)}_{1:n-1}, k)|x_{1:n})}{p(c^{(j)}_{1:n-1}|x_{1:n-1})}$$

(normalized such that all the new weights sum to 1). The sum in the numerator ensures that the new weight reflects the ability of this particle to predict the actual data point observed, rather than just how well the sampled component explains it (see Chen & Liu, 2000; Fearnhead, 2004, for further discussion).

One common problem with particle filters is that a small number of particles capture nearly all the weight, which drastically reduces the effective number of samples and hence the amount of information about the actual distribution they are approximating. In order to prevent this, when the variance of the weights becomes too high, a rejuvenation step resamples particles (with replacement) proportional to their weights, and resets the weights to be equal. The threshold for the variance of the weights was set to 50% of the mean of the weights (as suggested by Fearnhead, 2004).

Methods

In order to evaluate the particle filter as a model of phonetic distributional learning, I applied it to the problem of learning the English distinction between voiced /b/ and voiceless /p/, based on voice onset time (VOT). This is the primary acoustic cue to voicing for word-initial stops in English (Lisker & Abramson, 1964), and exhibits a clear bimodality. In order to simulate random VOT datasets, I fit two normal distributions to the VOTs for /b/ and /p/ from the Buckeye corpus (Pitt et al., 2007) by Nelson and Wedel (2017). This particular corpus shows low levels of talker variability in VOT (see Kleinschmidt, submitted). Fig. 1 shows an example randomly generated set of VOTs.

The Gibbs sampler and particle filter models were implemented in Julia (Bezanson, Edelman, Karpinski, & Shah, 2017).[2] Each was run on 200 randomly generated data sets of up to 10,000 observations, collecting intermediate results after 10, 100, and 1,000 observations. For comparison, 6-month-old infants in Bergelson and Aslin (2017) heard on average around 200 tokens of word-initial /b/ and /p/ in a single day, which would be 60,000 extrapolated to an entire year. For each dataset, the number of particles (or samples for the Gibbs sampler) varied (10, 100, and 1,000) and the concentration parameter α varied (0.01, 0.1, 1, and 10). The results for the Gibbs sampler were qualitatively the same across the number of samples and number of observations, so results are only shown for 1,000 samples (after 500 burnin) and 1,000 observations.

For the prior distribution over cluster means and variances, both used a weakly informative conjugate Normal-χ^{-2} prior (Gelman, Carlin, Stern, & Rubin, 2003), with μ_0 and σ^2_0 set to the overall mean and variance of all the VOTs, and $\kappa_0 = 0.05$ (to allow for significant possible variation in the cluster means) and $\nu_0 = 2$ (to constrain the variances to reasonable values).

[2] Available online at github.com/kleinschmidt/Particles.jl

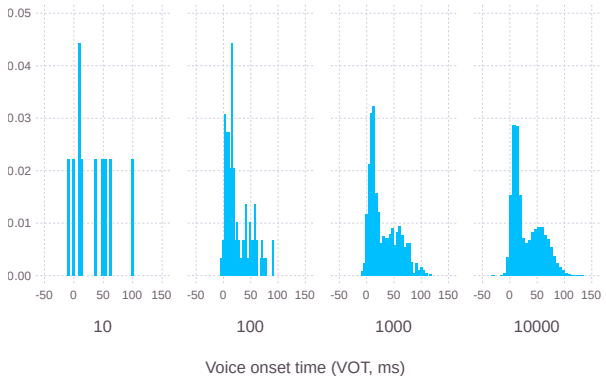

Figure 1: Example sample of VOTs used to assess distributional learning models. With fewer observations, the cluster structure is not obvious, but with more clusters it bnecomes clearer.

For each run, the approximated posterior distribution of cluster numbers is recorded. In the case of the Gibbs sampler, this is simply the proportion of samples with each number of clusters K. The calculation is analogous for the particle filter, except that the weights need to be taken into account. If particle j has $K^{(j)}$ clusters and weight $w^{(j)}$, then $p(K = k|x_{1:N}) = \sum_j w^{(j)} \delta_k(K^{(j)})$ (where $\delta_k(x)$ is the indicator function, which is 1 if $x = k$ and 0 otherwise).

Results

A natural measure of success is the probability assigned to $K = 2$, since the data was generated by a two-component mixture (Fig. 2). Additionally, Fig. 3 shows the expected number of clusters $E(K|x_{1:N}) = \sum_k k p(K = k|x_{1:N})$, and Fig. 4 shows the full distributions ($p(K|x_{1:N})$).

Gibbs sampler

Since the Gibbs sampler provides the "reference" approximation for these models, we first examine those results, shown as the gray lines in Fig. 2 and Fig. 3. Of the values tested here, the Gibbs sampler performs best with $\alpha = 0.1$, allocating the majority of the posterior probability to $K = 2$. Nevertheless, the Gibbs sampler maintains substantial uncertainty, allocating some 15-20%probability to clusterings with more or less than two clusters. For α an order of magnitude smaller or larger than this, the Gibbs sampler infers fewer or more (respectively) clusters than there actually are (Fig. 3), again with some substantial uncertainty.

Particle filter

The particle filter, unlike the Gibbs sampler, works best when $\alpha = 0.01$ (the smallest value tested). At this value, with a

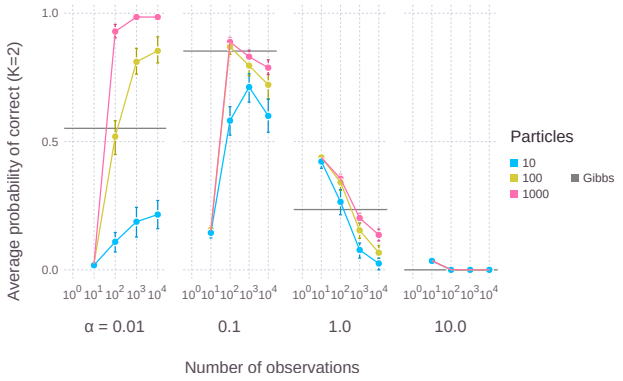

Figure 2: Posterior probability correct (two-cluster solutions) for particle filter (colors, 95% bootstrapped CIs over runs) and Gibbs sampler (gray horizontal lines). Gibbs sampler is shown for 1,000 observations and 1,000 samples.

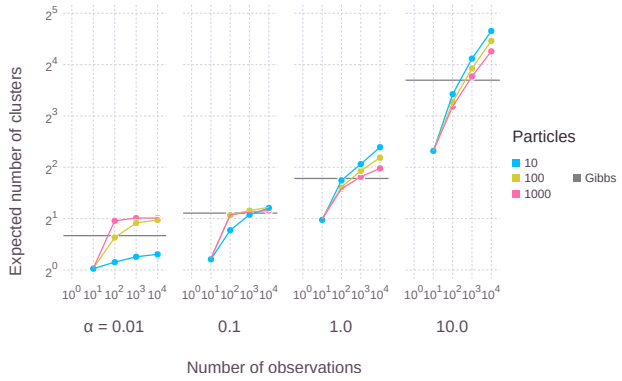

Figure 3: The expected number of clusters for particle filter (colors, 95% bootstrapped CIs over runs) and Gibbs sampler (gray horizontal lines). Gibbs sampler is shown for 1,000 observations and 1,000 samples

larger number particles, nearly 100% probability is assigned to the true $K = 2$ (first panel, Fig. 2).[3] For larger values of α, it eventually overshoots and infers more than 2 clusters. For instance, Fig. 2, second panel ($\alpha = 0.1$), shows that the probability assigned to $K = 2$ clusters rises to a maximum around 100 data points, but then falls with more data as more complex clusterings are increasingly preferred (Fig. 4). Even though the expected number of clusters is just slightly more than 2 (Fig. 3) as with the Gibbs sampler, there's no reason to think that it would continue to increase with more data, since there's no way for the particle filter to go back to simpler solutions once they've been forgotten.

Generally, particle filters with more particles better match the Gibbs sampler. This is not surprising: more particles mean more tolerance for uncertainty, which means that the particle filter is less committed to particular classifications that it made in the past. With few particles, it's possible (and indeed likely) that most of the particles agree on how the first data points should be classified, even if there is (ideally) some uncertainty there. Furthermore, with few particles, the prior has outsized influence. For low α, the 10-particle filter undershoots the number of clusters inferred by the Gibbs sampler (and by more particles), while for high α it overshoots.

Discussion

Rational approximations to rational models—like the particle filter explored here—provide a natural bridge between computational-level and algorithmic-level, cognitive models (Sanborn et al., 2010). These techniques are useful in part because they approximate optimal statistical inference as specified by the underlying Bayesian model. But they also allow us to explore—in a formal, quantitative way—how different kinds of cognitive constraints interact with the structure of the world to affect human cognition and learning. In the case of phonetic distributional learning, there is no shortage of cognitively plausible models (e.g., McMurray et al., 2009; Vallabha et al., 2007). The model explored here psychologically plausible—it takes one data point at a time and only maintains a small, finite number of hypothetical clusters–and also grounded in a computational-level model of the problem of distributional learning and its statistically optimal solution.

The approximation to the statistically optimal solution to distributional learning provided by the particle filtering algorithm of Chen and Liu (2000) can, in fact, recover the underlying structure of the particular model system I examined here (the American English /b/-/p/ contrast). More importantly, this particular sort of approximation constrains the model in a similar way that language learners are constrained: they cannot endlessly re-analyze every single sound they have ever heard, nor can they maintain an essentially infinite set of hypotheses about how those sounds should be clustered. In doing so, it provides some insight into how people actually solve the statistical problem posed by the computational-level model of distributional learning. Forgetting its own history actually *helps* the particle filter model when there is a strong bias towards fewer clusters (low α). The particle filter reliably arrives at the correct two-cluster structure for low values of α, even when the Gibbs sampler fails to do so, since the Gibbs sampler can continuously second-guess its decisions to create additional clusters.

The tendency for particle filters to get dug into a particular solution is often a shortcoming, but in this case it may be a benefit. It also suggests that such limited resources may actually be a benefit to human learners as well. It does not,

[3] Similar results are obtained for a three-component mixture, modeled after the languages described in Lisker and Abramson (1964) with three voicing categories. The primary difference is that more data is required to correctly infer that there are three categories for low values of α.

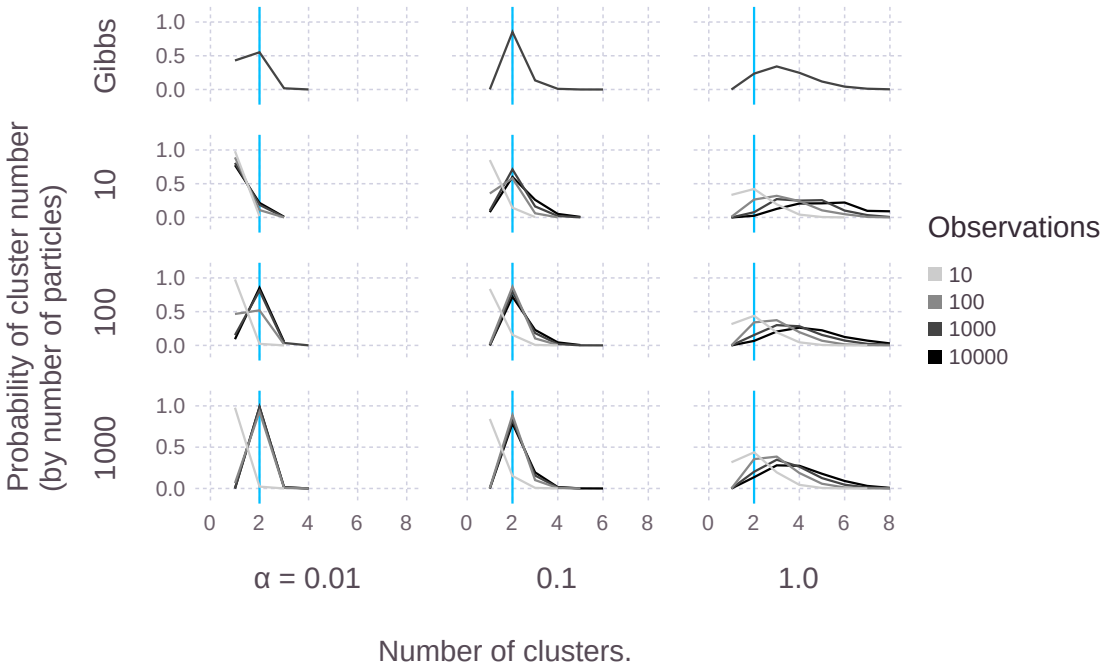

Figure 4: The posterior probability assigned to each possible number of components, as a function of the number of observations, number of particles, and α. Blue vertical lines show the true number of clusters $K = 2$.

however, require that such limitations *change* with development (as in Newport, 1990) or require or particularly benefit from a specific input (as in Elman, 1993). In fact, the constraints imposed by limited tolerance for uncertainty in the particle filter model examined here are in a sense the *opposite* of the "less is more" hypothesis: these constraints *allow* the learner to impose stronger a priori preference for simple explanations and still learn, rather than *constituting* a simplicity preference in and of themselves. Indeed, one of the advantages of the kind of cognitively constrained rational models described here is that it allows a principled exploration of the way that cognitive constraints interact with different kinds of input and structural assumptions on the part of the learner (Rohde & Plaut, 1999; Siegelman & Arnon, 2015).

Additionally, specifying and exploring such models of human behavior also has the potential to improve basic computational techniques as well. We know that humans do manage to learn the underlying cluster structure from unsupervised input like this. The particular ways in which *models* fail at this task can be instructive for creating better models. This may mean taking into account higher-order structure where it's present, like phonotactic and lexical regularities (Feldman et al., 2013). But it might also motivate specific techniques to get around the difficulty of moving from more complex to less complex clusterings, like adding the possibility of merging clusters when rejuvenating particles (analogously to reversible-jump MCMC, Green, 1995).

Finally, distributional learning continues even in adult listeners, both in second language learning (e.g., Pajak & Levy, 2011) and in adapting to unfamiliar talkers (Clayards, Tanenhaus, Aslin, & Jacobs, 2008; Xie, Theodore, & Myers, 2017). The same Bayesian models that have been applied to distributional learning in acquisition can explain many of the patterns observed in adaptation (Kleinschmidt & Jaeger, 2015). Listeners are able to maintain some uncertainty about how previous tokens ought to be categorized (Bicknell, Tanenhaus, & Jaeger, submitted), but it is currently unclear how this constrains their ability to learn from distributional information. The models explored here could just as easily be applied to adaptation in adults as acquisition in children.

Conclusion

Approximately-rational models like the particle filter provide a possible bridge between computational-level models and psychologically plausible algorithmic-level cognitive models. A particle filter model of phonetic distributional learning is able to learn the underlying cluster structure of the English /b/-/p/ contrast based only on the distribution of a single cue (VOT). This shows that it is possible to approximate optimal Bayesian inference in this domain without making the psychologically-implausible assumptions of batch processing and unlimited tolerance of uncertainty. However, the behavior of this approximation also diverges in potentially interesting ways from a less-constrained approximate infer-

ence model (a Gibbs sampler), suggesting that the constraints posed by limited cognitive resources are a critical piece of the puzzle in understanding cognition, even for computational-level modeling (Marr, 1982).

References

Bergelson, E. & Aslin, R. N. (2017). Nature and origins of the lexicon in 6-mo-olds. *Proceedings of the National Academy of Sciences*, 201712966. doi:10 . 1073 / pnas . 1712966114. pmid: 29158399

Bezanson, J., Edelman, A., Karpinski, S., & Shah, V. (2017). Julia: A Fresh Approach to Numerical Computing. *SIAM Review*, *59*(1), 65–98. doi:10.1137/141000671

Bicknell, K., Tanenhaus, M. K., & Jaeger, T. F. (submitted). Listeners can maintain and rationally update uncertainty about prior words.

Chen, R. & Liu, J. S. (2000). Mixture Kalman Filters. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, *62*(3), 493–508. JSTOR: 2680693

Clayards, M. A., Tanenhaus, M. K., Aslin, R. N., & Jacobs, R. a. (2008). Perception of speech reflects optimal use of probabilistic speech cues. *Cognition*, *108*(3), 804–9. doi:10.1016/j.cognition.2008.04.004

Elman, J. L. (1993). Learning and development in neural networks: The importance of starting small. *Cognition*, *48*(1), 71–99. doi:10.1016/0010-0277(93)90058-4

Fearnhead, P. (2004). Particle filters for mixture models with an unknown number of components. *Statistics and Computing*, *14*(1), 11–21. doi:10.1023/B:STCO.0000009418.04621.cd

Feldman, N. H., Griffiths, T. L., Goldwater, S., & Morgan, J. L. (2013). A role for the developing lexicon in phonetic category acquisition. *Psychological Review*, *120*(4), 751–778. doi:10.1037/a0034245

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2003). *Bayesian Data Analysis* (Second). Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis.

Gershman, S. J. & Blei, D. M. (2012). A tutorial on Bayesian nonparametric models. *Journal of Mathematical Psychology*, *56*(1), 1–12. doi:10.1016/j.jmp.2011.08.004

Goldwater, S., Griffiths, T. L., & Johnson, M. (2009). A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, *112*(1), 21–54. doi:10.1016/j.cognition.2009.03.008. PMID: 19409539

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, *82*(4), 711–732. doi:10.1093/biomet/82.4.711

Kleinschmidt, D. F. (submitted). Structure in talker variability: How much is there and how much can it help? doi:10.17605/OSF.IO/A4TKN

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

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

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

Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, *82*(3), B101–11. PMID: 11747867

McMurray, B., Aslin, R. N., & Toscano, J. C. (2009). Statistical learning of phonetic categories: Insights from a computational approach. *Developmental Science*, *12*(3), 369–378. doi:10.1111/j.1467-7687.2009.00822.x

Nelson, N. R. & Wedel, A. (2017). The phonetic specificity of competition: Contrastive hyperarticulation of voice onset time in conversational English. *Journal of Phonetics*, *e-pub*, 1–20. doi:10.1016/j.wocn.2017.01.008

Newport, E. L. (1990). Maturational Constraints on Language Learning. *Cognitive Science*, *14*(1), 11–28. doi:10.1207/s15516709cog1401_2

Pajak, B. & Levy, R. (2011). Phonological Generalization from Distributional Evidence. In L. Carlson, C. Hoelscher, & T. F. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 2673–8). Cognitive Science Society.

Pitt, M. A., Dilley, L. C., Johnson, K., Kiesling, S., Raymond, W., Hume, E., & Fosler-Lussier, E. (2007). Buckeye Corpus of Conversational Speech (2nd release). Columbus, OH: Department of Psychlogy, Ohio State University. Retrieved from www.buckeyecorpus.osu.edu

Rohde, D. L. T. & Plaut, D. C. (1999). Language acquisition in the absence of explicit negative evidence: How important is starting small? *Cognition*, *72*(1), 67–109. doi:10.1016/S0010-0277(99)00031-1

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2010). Rational approximations to rational models: Alternative algorithms for category learning. *Psychological Review*, *117*(4), 1144–67. doi:10 . 1037 / a0020511. PMID: 21038975

Siegelman, N. & Arnon, I. (2015). The advantage of starting big: Learning from unsegmented input facilitates mastery of grammatical gender in an artificial language. *Journal of Memory and Language*, *85*, 60–75. doi:10 . 1016 / j . jml . 2015.07.003

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

Xie, X., Theodore, R. M., & Myers, E. B. (2017). More than a boundary shift: Perceptual adaptation to foreign-accented speech reshapes the internal structure of phonetic categories. *Journal of Experimental Psychology: Human Perception and Performance*, *43*(1), 206–217. doi:10 . 1037 / xhp0000285