

Modeling Unsupervised Perceptual Category Learning

Brenden M. Lake, Gautam K. Vallabha, and James L. McClelland

Department of Psychology, Stanford University

Jordan Hall Building 420, Stanford, CA 94305

Email: {brenden, mcclelland}@stanford.edu

Abstract—During the learning of speech sounds and other perceptual categories, category labels are not provided, the number of categories is unknown, and the stimuli are encountered sequentially. These constraints provide a challenge for models, but they have been recently addressed in the Online Mixture Estimation model of unsupervised vowel category learning [1]. The model treats categories as Gaussian distributions, proposing both the number and parameters of the categories. While the model has been shown to successfully learn vowel categories, it has not been evaluated as a model of the learning process. We account for three results regarding the learning process: infants’ discrimination of speech sounds is better after exposure to a bimodal rather than unimodal distribution [2], infants’ discrimination of vowels is affected by acoustic distance [3], and subjects place category centers near frequent stimuli in an unsupervised visual classification task [4].

I. INTRODUCTION

The ability to categorize objects is critical for perception. Knowing an object is in the category “chicken” provides crucial information about that object – such as it has feathers, it is edible, and it can fly. Much modeling work has investigated how categories are learned, e.g. [5]–[7].

While category learning is often facilitated by associating objects with category labels, categories can be acquired by mere exposure to stimuli – no labels included. For instance, during the first year of life, infants begin acquiring the speech sound categories of their native language; sensitivity to non-native contrasts decreases [8] and sensitivity to native contrasts increases [9]. In the visual modality, Rosenthal et al. [4] found that subjects’ categorical decisions, without feedback, were influenced by the distributional properties of the stimuli.

How can category structure be learned without labels? Models face two challenges: (i) the number of categories to learn is unknown, and (ii) the stimuli are encountered one by one in mixed order instead of all at once [1]. There has been some recent progress addressing these issues in models of speech category learning [1], [10], [11]. The Online Mixture Estimation (OME) algorithm [1] is an online variant of Expectation-Maximization (EM) that addresses both of the above issues. It tries to find a set of Gaussian categories that account for a sequence of stimuli, proposing both the number and parameters of the categories. The model is also somewhat biologically plausible since a topographical network can serve as an approximation. In Vallabha et al. [1], the OME algorithm successfully learned the number and parameters of

multidimensional vowel categories in English and Japanese [1]. However, successfully learning vowel categories does not imply the algorithm successfully models the *learning process*. The current work provides such an evaluation.

What quantitative means are available to assess the learning process? Category learning is often marked by changes in discrimination. During the learning process, there is evidence for improving discrimination across category boundaries (*acquired distinctiveness*) [9], [12], [13] and declining discrimination within category boundaries (*acquired similarity*) [8], [14], terms from [12]. In a very similar algorithm to OME but restricted to one dimensional stimuli, McMurray et al. [11] found both effects, defining discrimination as the extent to which two stimuli are members of different estimated categories. Should OME show similar effects, it would capture two important aspects of the category learning process.

We apply the OME algorithm to three other results regarding the learning process. The first model is of Maye et al.’s [2] study where infants are sensitized to a [da]-[ta] continuum of speech sounds with either a bimodal or unimodal distribution. Infants sensitized to the bimodal distribution showed better discrimination of the endpoints. Second, we investigated how discrimination might develop over time. In Sabourin et al. [3], infants showed better discrimination of acoustically more distinct vowels, and we applied the OME model to this vowel space. Third, we extended the model from the domain of speech to vision, modeling Rosenthal et al.’s [4] unsupervised category learning task where subjects’ categorical choices were influenced by the stimulus distribution.

II. THE ONLINE MIXTURE ESTIMATION MODEL

The Online Mixture Estimation (OME) algorithm treats categories as multivariate Gaussians distributions and gradually estimates the category structure from a sequence of stimuli, proposing both the number of categories and their parameters (Figure 1). The model begins with many (50 or more) initial *guess categories* distributed randomly about the space in which the data reside.¹ Each guess category has an associated *mixing probability*, the probability of the guess category contributing a random token to the stimulus set. In our simulations, all guess categories were initialized to be equally likely. For speed

¹In our simulations, the initial guess category covariance matrices were diagonal (although the full covariance matrix is updated during learning). The initial variances along the diagonal were random.

we tend to eliminate them when their mixing probability falls below a threshold but this is not necessary.

During learning, the probability that category c is responsible for stimulus S is an application of Bayes' Rule,

$$p(c|S) = \frac{p(S|c)p(c)}{\sum_c p(S|c)p(c)} \quad (1)$$

where $p(S|c)$ is the Gaussian density for category c , and $p(c)$ is the mixing probability. We refer to the posterior probability $p(c|S)$ as the *responsibility* of category c for stimulus S .

Learning occurs online as stimuli are presented one at a time in random order. For the presentation of stimulus S , the responsibility $p(c|S)$ is first calculated for each category (Equation 1). The mean and covariance matrix for each category are updated to account better for S , with the update size proportional to the responsibility and a learning rate. Then, the mixing probability of the winning category (the category with highest responsibility) is increased such that the mixing probabilities still sum to 1 across categories. For a precise formulation of the algorithm, see [1] Methods section.

We extend the OME model to account for discrimination behaviors, defining pairwise discrimination as

$$Discrimination(S_a, S_b) = \sqrt{\sum_c (p(c|S_a) - p(c|S_b))^2} \quad (2)$$

which is the Euclidean distance between the responsibility vectors for two stimuli S_a and S_b (a similar approach was taken in McMurray et al. [11] with root mean square distance). If two stimuli are likely to be categorized as the same, they are hard to discriminate. If two stimuli are likely to be categorized as different, they are easy to discriminate. Thus, discrimination is defined as a function of categorization, determined by the current category representations during learning.

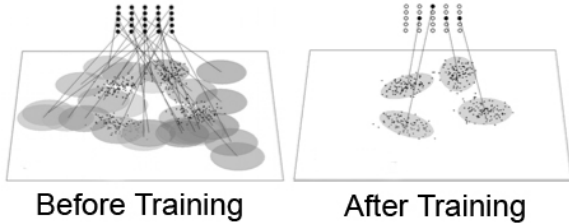


Fig. 1. This figure illustrates OME learning stimuli in 2D space. OME is initialized with many guess categories spread over input space (grey ovals), with equal mixing probabilities (grid of black circles). Learning involves presenting the stimuli (black dots) one by one and updating the categories. After learning, unneeded guess categories have mixing probabilities near zero (open circles in grid), and OME has fit the remaining guess categories to the stimulus clusters. Figure reprinted with permission from [1], Copyright (2007) National Academy of Sciences, U.S.A.

III. EXPERIMENT 1: BETTER DISCRIMINATION WITH BIMODAL RATHER THAN UNIMODAL DISTRIBUTION

We model a study from Maye et al. [2] that investigated a mechanism of developmental change in speech perception. The authors sensitized six and eight-month-old infants to a

continuum of unaspirated coronal stops ranging from a [da]-like sound to a more [ta]-like sound, which is a contrast infants of this age have been found to discriminate without sensitization [15]. This continuum was drawn from either a unimodal or bimodal distribution (Figure 2). As a result of sensitization, the authors predicted that “infants exposed to a bimodal distribution would form a two-category representation of these sounds, while infants exposed to a unimodal distribution would form a one-category representation.” As a consequence, the bimodal distribution would facilitate better discrimination of the endpoints, a conclusion supported by the data [2].

Since this has been suggested as a mechanism [2] for the developmental changes seen during infant speech category acquisition [8], [9], it is crucial that OME shows a similar effect – gravitating towards a structure in the bimodal case that supports better discrimination. In this experiment, we investigate how OME reacts to unimodal vs. bimodal stimuli.

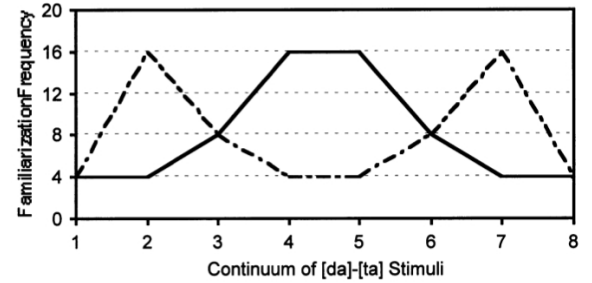


Fig. 2. Bimodal distribution (dotted) and unimodal distribution (solid) of [da]-[ta] stimuli. The x-axis is the discrete stimuli from the most [da]-like (left) to [ta]-like (right). The y-axis is frequency of presentation for the infants and model. Reprinted from [2] Copyright (2002), with permission from Elsevier.

A. Model

We used the exact bimodal and unimodal distributions from Maye et al. [2]. Thus, both the infants and model were presented with 64 stimulus tokens. The stimuli presented to the model were simply the values 1 through 8, presented in random order. For the model parameters, there were 50 initial guess categories, with means randomly chosen from a uniform distribution from values -1 to 10 and standard deviations from 2 to 4. Thus, the algorithm began with a range of initial categories with different standard deviations and means, disallowing it to anticipate the structure of the actual data. It seems reasonable that an organism would have initial categories in a variety of sizes, but we do not have any specific reasons for selecting one range of standard deviations over another. The learning rate was .05, and guess categories were removed after dropping below mixing probability .001. Using the discrimination metric from Equation 2, we measured discrimination of the [da]-[ta] continuum endpoints as training progressed. Thus, we determined if an underlying bimodal distribution leads to better endpoint discrimination than a unimodal distribution. We ran 24 replications of each distribution with random initial categories and stimulus presentation order.

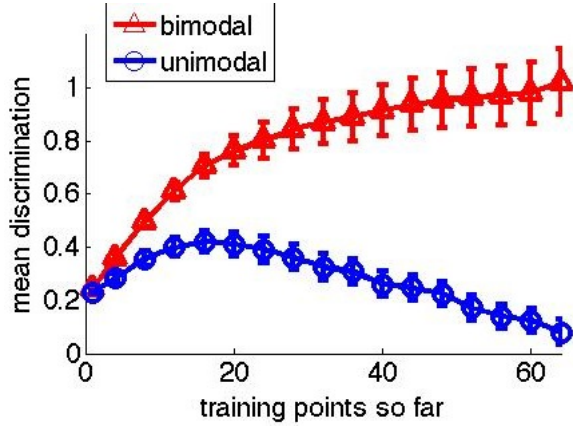


Fig. 3. Time course of [da]-[ta] discrimination for the two distributions averaged over runs. Error bars are standard error.

B. Results

We found significantly better endpoint discrimination in models exposed to the bimodal distribution, as was the case for the infants in [2]. After learning, the mean discrimination was 1.02 for the bimodal models and 0.079 for the unimodal models (independent samples $t(46) = 7.12, p < .001$). The time course of discrimination is plotted in Figure 3. There was a general increase in discrimination for the bimodal distribution. Despite an initial increase in discrimination for the unimodal distribution, the models showed an overall decrease in discrimination (paired $t(23) = 3.27, p < .01$).

The model accounts for two aspects of the data. First, the OME model shows superior discrimination after exposure to a bimodal rather than unimodal distribution. Second, the infants who were familiarized to the unimodal distribution did not significantly discriminate the endpoints in Maye et al. [2], although a past study has shown infants of this age can make this discrimination [15]. Thus, the unimodal distribution likely caused a reduction in discrimination. The OME model accounts for this reduction. If this sensitivity to unimodal vs. bimodal contrasts is a central mechanism in infant speech category acquisition, OME provides an account of how the learning process could occur – through small, online updates to the category structure as the infant receives speech tokens.

IV. EXPERIMENT 2: CLOSE CATEGORIES ACQUIRE DISTINCTIVENESS SLOWER

We have shown that the OME model accounts for an important proposed mechanism of speech acquisition. Furthermore, as mentioned, the OME algorithm has found success in learning vowel categories [1]. Can we say more about how discrimination and category structure develops throughout the learning process? During infant speech acquisition, speech tokens in the same category become harder to discriminate (acquired similarity) and tokens in different speech categories become easier to discriminate (acquired distinctiveness). McMurray et al. [11] found both effects with a similar model and discrimination metric. Does OME show these effects?

Furthermore, in Sabourin et al. [3], infants showed superior discrimination on an acoustically more distinct contrast. If two speech tokens are from different categories, are they easier to discriminate if they belong to further apart categories?

In this experiment, we model Sabourin et al. [3] who tested 8-month-old English monolinguals on their ability to discriminate acoustically similar vowels /e/ vs. /I/ and /e/ vs. /E/ (Figure 4). Behavioral tests indicated that infants could discriminate the acoustically more distinct /e/ and /E/ but not /e/ and /I/. Further investigation using event-related potentials found that infants could discriminate both contrasts, but it also showed infants had more difficulty discriminating /e/ vs. /I/. In this experiment, we trained the model on points drawn from these three vowel categories (Figure 4), and tracked discrimination between the categories. The stimuli were 2D in F1 and F2 space. We would expect discrimination to increase faster between /e/ vs. /E/ rather than /e/ vs. /I/. If the initial guess categories are wide enough to spread across two vowels, they might provide similar responsibilities to tokens of neighboring vowels, contributing little to discrimination. However, these wide categories may aid discrimination between further apart categories. The OME model was trained 10 different times with different draws from the vowel stimuli (Figure 4). See footnote for details.²

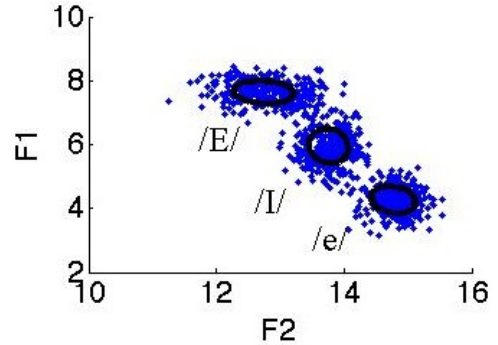


Fig. 4. Vowel categories estimated from Sabourin et al. [3] and converted to the Bark scale. The 1000 blue points are a random draw, and the black circles are the categories OME found, plotted 1 s.d. along each principal axis.

A. Results

To calculate discrimination between two categories rather than two stimuli as in Equation 2, we simply drew 50 additional test points from each vowel before training. Then we define the *perceptual distance* between two categories as the mean pairwise discrimination between the test points of those categories (with the first test point from each category paired,

²We drew 1000 points from these three vowel, assumed to be Gaussian, with equal probability from each. The points were then converted to the Bark scale (1 to 24, corresponding to the first 24 critical bands of hearing). There were 50 initial guess categories. We ran the algorithm 10 times with different draws for the vowel points and initial categories. The authors provide the means but not the standard deviations for the vowel categories, so they were estimated as 1/3 the range along F1 and F2 with no covariance. For initializing the guess categories, the following were randomly drawn from uniform distributions in Bark: F1 means from 3.9 to 8, F1 s.d. from .5 to 1, F2 means from 12.4 to 15, and F2 s.d. from .5 to 1.5. The learning rate was .01.

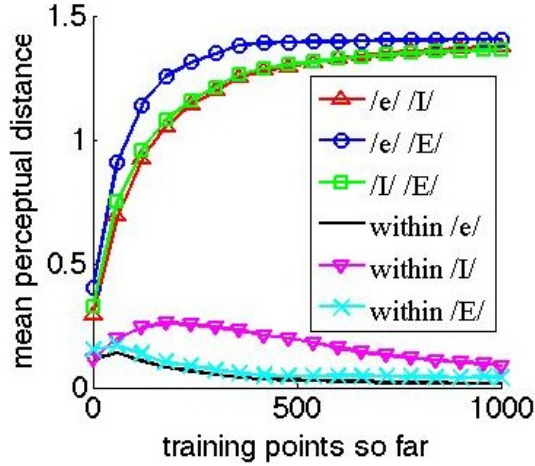


Fig. 5. This figure plots perceptual distance across training, averaged over 10 runs. The largest standard error, if shown, would be 0.036. In the legend, the first three lines illustrate increasing discrimination between vowels, and the last three lines illustrate decreasing discrimination within vowels.

the second paired, and so on). We use the analogous method of pairing test points within a category to measure perceptual distance within a category. Thus, *acquired distinctiveness* is increasing perceptual distance between categories, and *acquired similarity* is decreasing perceptual distance within categories.

OME learned 3 vowel categories in all 10 runs. The model shows acquired distinctiveness [9] and acquired similarity [8] (Figure 5). We also found that the model’s discrimination of vowels is affected by acoustic distance (Figure 5). In particular, the perceptual distance between vowels /e/ vs. /E/ grew significantly faster than /e/ vs. /I/, accounting for the result in Sabourin et al. [3].³ OME thus predicts that further apart categories differentiate faster, and accounts for acquired distinctiveness and similarity, found in both supervised and unsupervised category learning in various modalities.

V. EXPERIMENT 3: PLACING CATEGORIES IN FREQUENT STIMULUS AREAS

While both previous experiments were auditory, the same principles of unsupervised learning, online updating, and unknown number of categories the OME algorithm instantiates can be applied to visual category learning. In the next experiment, we modeled Rosenthal et al.’s [4] unsupervised categorization task, where subjects categorized simple, one-dimensional visual stimuli. Subjects saw a sequence of vertical stripes with varying width, and they were informed “only that they would see stimuli of one or more kinds and should classify them accordingly,” with eight keys available for responses. The stimulus width was drawn from a frequency distribution with several Gaussian peaks (either 3-

peaks, 4-peaks, or uniform distribution). While subjects’ post-test frequency evaluations did not match the actual frequency, their categorical decisions were nonetheless influenced by the stimulus frequency; subjects placed category centers around stripe widths that appeared most frequently.

To account for this implicit effect, Rosenthal et al. proposed a self-organizing neural network model utilizing Hebbian learning. While the neural network model accounts for the gradual organization of category structure, it makes one wrong prediction using settling time as a measure of reaction time (RT). Subjects showed lower RT for extreme stimuli (extremely narrow or wide) than for peak stimuli, while the model showed the opposite. In this experiment, we show that the OME algorithm provides an account for the data in [4].

A. Model

OME was trained for 4,096 trials, as were subjects. There were 20 replications for each condition (3-peak, 4-peak, and uniform) with different random stimulus sequences and starting guess categories. Rosenthal et al. [4] allowed the stripe width to vary from 1-512 pixels, divided into 36 sample bins. The model stimuli were integers from 1 to 36 drawn from the same distributions, but the stimuli were perturbed by Gaussian noise with standard deviation 2 before training OME. There were 100 initial guess categories with means picked from a uniform distribution from 1 to 36 and standard deviations were picked from a uniform distribution from $(1/2)\sigma$ to 25σ for the 3-peak σ . The learning rates were .0025 and .005 for mixing probability and category updates, respectively. Guess categories below mixing probability .0001 were removed.

B. Results

Both the OME model and human subjects organized their category structures based on the distributional properties of the stimuli. During the fourth session (last quarter of training), the subjects and model showed a clear tendency to organize category centers near peak frequencies and boundaries between peak frequencies (see Figure 6A). See footnote for details on calculating the centers and boundaries.⁴

Rosenthal et al. [4] also investigated the time course of categorization. In Figure 6B, the histogram displayed in Figure 6A was calculated for each of the four sessions and averaged across multiple peaks. Thus, it illustrates how learning evolves across sessions. As with the subjects, the model’s centers and boundaries are increasingly influenced by the stimulus distribution as the sessions progress. However, evolution of

³To check the significance of the faster growth, we fit a damped exponential curve to each distance trajectory separately for the 10 runs: $f(x) = c(1 - e^{-dx})$. Parameter d corresponds roughly to “rate of increase,” where a larger value is a faster increase. Thus, we take d as an approximation for how fast two vowels differentiate. By this measure, /e/ vs. /E/ differentiates significantly faster (paired $t(9) = 7.39, p < .001$).

⁴The calculation of centers and boundaries followed [4]. First, the authors defined a sorting coherence function, for a particular subject, $\chi(b, c)$ as the fraction of presentations a stimulus in bin b was classified as class c . To simulate this measure in our model, we defined a classification for a particular stimulus as the guess category with the highest responsibility (Equation 1). Then, calculating the sorting coherence function for a session is straightforward. Second, the center and boundaries must be calculated for each category c chosen for classification in a particular session. The center was defined as $\sum_{b=\text{left boundary}}^{\text{right boundary}} b \cdot \chi(b, c) / \sum_{b=\text{left boundary}}^{\text{right boundary}} \chi(b, c)$. The location of the left (right) boundary was the last bin in which $\chi(b, c) \leq 0.5$, when starting at the leftmost (rightmost) bin and moving towards the right (left). The edge bins were the boundaries for the edge categories [16].

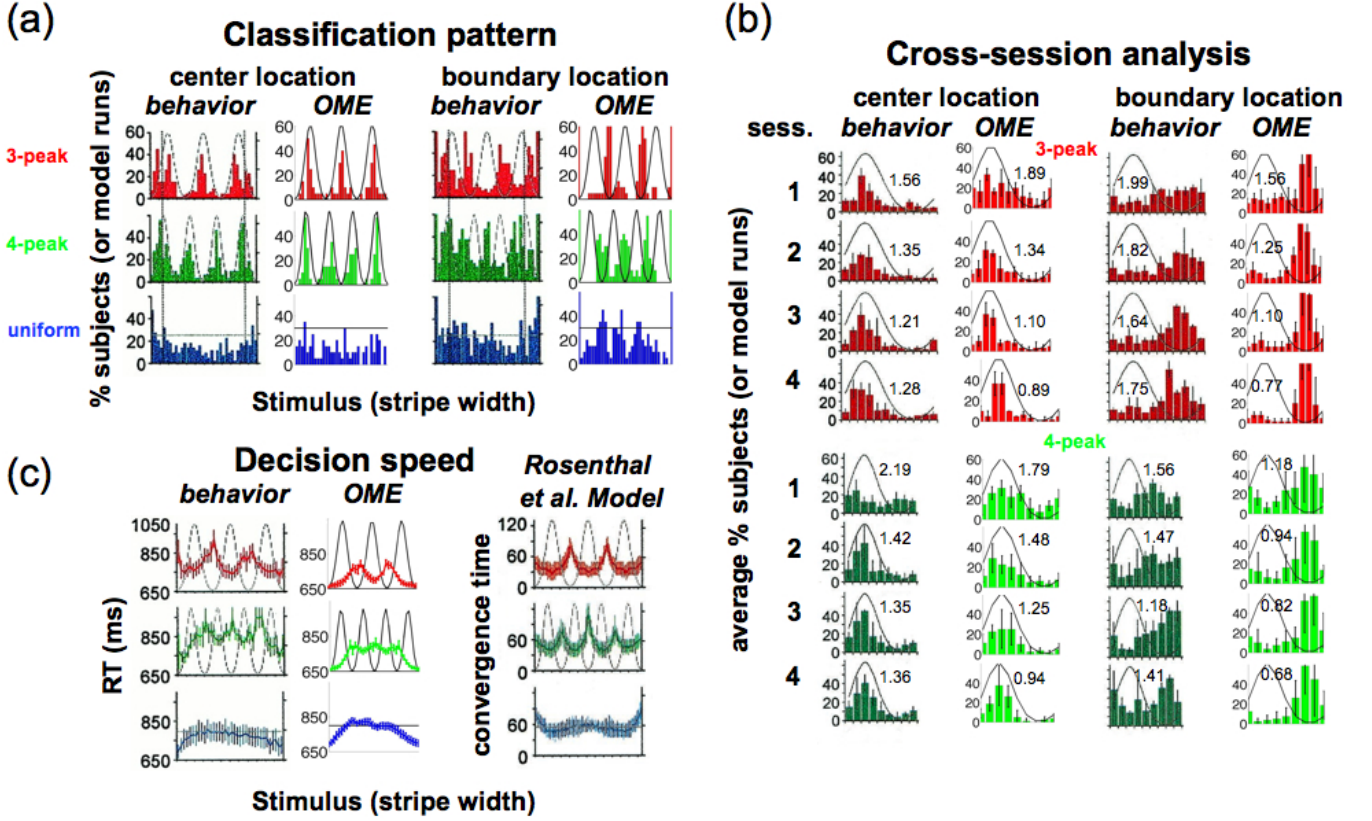


Fig. 6. All of the diagrams under headings “behavior” and “Rosenthal et al. Model” are reprinted with permission from [4], Copyright (2001) National Academy of Sciences, U.S.A. Note that some bars in this figure extend beyond 60%. (a) For the fourth session, this figure shows the histogram of center and boundary locations for the subjects and the model. The bars on the histogram represent the percent of subjects (or model runs in the case of OME) to have a center or boundary in that stimulus bin. The x-axis denotes the width of the stripes, with the stimulus frequency distribution illustrated by the black curve. (b) For all four sessions, this figure shows the histogram of center and boundary locations, averaged across peaks. Thus, the bars represent average percent of subjects having a center/boundary within a certain distance from a distribution peak. The numbers above the diagrams are a measure of accuracy, with lower numbers indicating a closer match of the centers/boundaries to the stimulus distribution centers/boundaries. They are calculated as: $\sum_{k=1}^{nbars} (height_k) * (number\ of\ bars\ from\ bar_k\ to\ distribution\ center\ [or\ boundary]) / \sum_{k=1}^{nbars} height_k$ where $nbars$ is the number of bars. The heights from the behavioral data are estimated from the graphs in [4]. Error bars are s.d. (c) For the fourth session, this figure shows RT for the behavior and OME, averaged across subjects (or runs). The right column shows convergence time of the Rosenthal et al. model, which increases at the edges where the behavior and OME do not. The error bars are s.e. for the behavior/OME and s.d. for the Rosenthal et al. model.

the boundaries is faster in the model than in subjects. Also, subjects’ categories evolve more slowly in the four-peak case than the three-peak case, but not in the model.

Also, Rosenthal et al. [4] measured subject response time (RT). To simulate RT in the model for a stimulus, we took the largest guess category responsibility (Equation 1) minus the second largest, which is inversely related to RT and scaled into the subject range.⁵ Thus, if one guess category was clearly the most responsible for a stimulus, the responsibility difference is large and RT is small. If two guess categories were similarly responsible, the responsibility difference is small and the RT is large. In Figure 6C, we compare RT in the subjects and the model for the fourth session, finding a clear tendency for longer RT near the middle of two peaks.

Interestingly, Rosenthal et al. [4] found that stimuli near the

edges of the range were categorized faster (particularly for the 4-peak and uniform cases). Their neural network model makes the opposite prediction using settling time as a measure of RT, while OME makes the correct prediction (Figure 6C). In the OME model, the category closest to an edge, for example the left edge, would be mostly responsible for edge stimuli to the left of the category. Therefore, the category likely does not have much competition for the edge stimuli, producing a small second largest category responsibility and thus a low simulated RT. Thus, the model can account for this edge effect.

This model is robust across a range of parameters, but it is interesting to examine values that do not fit the data. Initializing OME with many small guess categories, such as s.d. $(1/2)\sigma$ where σ is the 3-peak Gaussian s.d., without an increase in learning rate results in too many categories in early learning to match the subject histograms. Furthermore, the model can break down by starting with only large guess categories, such as s.d. 6σ , where it forms too few categories.

Given the similarity of performance between OME and the

⁵We scaled the difference in responsibility into simulated RT in ms: $RT = -400(Largest\ responsibility - 2nd\ Largest\ Responsibility) + 1050$. The difference in responsibilities could vary from 0 to 1, and thus simulated RT can vary from 650 to 1050 ms.

behavior, OME is a particularly good model of unsupervised categorization when stimuli are drawn from a Gaussian mixture. As with the subjects, the inferred center and boundary locations were influenced by the distribution frequencies, with the influence evolving over training time. Furthermore, as with the subjects, it seems natural that OME would be more certain about a categorization query for peak and edge stimuli.

VI. GENERAL DISCUSSION

Categorization is essential to perception, and much of category learning is unsupervised. How can category structure be learned from just a sequence of stimuli? The Online Mixture Estimation (OME) algorithm [1] has provided some progress, showing that the number and parameters of vowel categories can be learned through online updating. However, showing the algorithm can solve the required learning problem [1] does not show the algorithm is a model of the processes to get there.

From this work, there are several results to recommend OME as a process model of category learning. In Experiment 1, the model produced better discrimination after exposure to a bimodal rather than unimodal distribution, accounting for a proposed mechanism of infant speech acquisition [2]. To investigate how discrimination develops over time, in Experiment 2, the OME model was applied to a crowded vowel space [3]. Both infants and the model showed better discrimination of a more acoustically dissimilar contrast than a more similar contrast. Also in Experiment 2, discrimination between vowels increased over time (acquired distinctiveness) and discrimination within vowels decreased over time (acquired similarity); both effects have empirical evidence from various modalities [8], [9], [12]–[14] and follow naturally from the modeling framework. In Experiment 3, the OME model showed that the same principles governing auditory category learning can be applied to visual category learning, where both subject and model categorization choices and response times were influenced by the distribution of stimuli. As previously noted, the model does not yet match some aspects of the human time-evolution data, an issue we are currently investigating.

More generally, OME provides an elegant solution to the problems of (1) scalability, (2) sensitivity, (3) revisability, and (4) cross-modal fusion in category learning. (1) Regarding scalability, OME's computational complexity is largely independent of the number of functionally useful data categories, only influencing complexity by affecting the number of guess categories. (2) Also, OME is sensitive to overlapping categories, reconstructing the data distribution from categories with means as close as 2.5 s.d. apart. (3) Additionally, the model is able to revise its solution if presented dynamic data categories. If a data category is removed from presentation, OME's corresponding guess category will progressively drop in mixing probability. If a new data category is added during learning and unused guess categories are not removed, they can provide a mechanism for adding categories. (4) Furthermore, OME can learn cross-modal categories. Combining auditory and visual dimensions, such as speech and the speaker's mouth position, is entirely compatible with the approach.

There is much to explore in future work, including how repeated stimulation affects sensitivity. In Jenkins et al. [17], monkeys placed their fingers in contact with a rotating disk in exchange for reward many times a day over months. This repeated stimulation of the fingertips resulted in shrinkage of receptive fields and expanded cortical area for the stimulated surface, likely improving sensitivity in this region. In contrast, repeated and concentrated stimulation in OME would likely form a category, resulting in decreased sensitivity due to acquired similarity. The issue here is empirical as well as theoretical; it is not yet clear why some experiments produce increased sensitivity, while others produce decreased sensitivity, to clustered stimuli. We are examining whether modifications to OME could produce the opposite behavior, potentially providing insight into this deep question.

REFERENCES

- [1] G. K. Vallabha, J. L. McClelland, F. Pons, J. F. Werker, and S. Amano, "Unsupervised learning of vowel categories from infant-directed speech," *Proceedings of the National Academy of Science*, vol. 104, no. 33, pp. 13 273–13 278, 2007.
- [2] J. Maye, J. F. Werker, and L. Gerken, "Infant sensitivity to distributional information can affect phonetic discrimination," *Cognition*, vol. 82, pp. B101–B111, 2002.
- [3] L. Sabourin, J. F. Werker, L. Bosch, and N. Sebastián-Gallés, "Perceiving vowels in a tight vowel space: evidence from monolingual infants," *Developmental Science*, under revision.
- [4] O. Rosenthal, S. Fusi, and S. Hochstein, "Forming classes by stimulus frequency: behavior and theory," *Proceedings of the National Academy of Science*, vol. 98, pp. 4265–4270, 2001.
- [5] T. T. Rogers and J. L. McClelland, *Semantic Cognition: A Parallel Distributed Processing Approach*. Cambridge, MA: MIT Press, 2004.
- [6] J. R. Anderson, "The adaptive nature of human categorization," *Psychological Review*, vol. 98, pp. 409–429, 1991.
- [7] T. L. Griffiths, A. N. Sanborn, K. R. Canini, and D. J. Navarro, "Categorization as nonparametric bayesian density estimation," in *The probabilistic mind: Prospects for rational models of cognition*, M. Oaksford and N. Chater, Eds. Oxford: Oxford University Press, 2008.
- [8] J. F. Werker and R. C. Tees, "Cross-language speech perception: evidence for perceptual reorganization during the first year of life," *Infant Behavior and Development*, vol. 7, pp. 49–63, 1984.
- [9] P. K. Kuhl, E. Stevens, A. Hayashi, T. Deguchi, S. Kiritani, and P. Iverson, "Infants show a facilitation effect for native language phonetic perception between 6 and 12 months," *Developmental Science*, vol. 9, no. 2, pp. F13–F21, 2006.
- [10] B. de Boer and P. K. Kuhl, "Investigating the role of infant-directed speech with a computer model," *Acoustics Research Letters Online*, vol. 4, pp. 129–134, 2003.
- [11] B. McMurray, R. N. Aslin, and J. Toscano, "Statistical learning of phonetic categories: Insights from a computational approach," *Developmental Science*, in press.
- [12] R. K. Goldstone, "Influences of categorization on perceptual discrimination," *Journal of Experimental Psychology: General*, vol. 123, no. 2, pp. 178–200, 1994.
- [13] S. V. Stevenage, "Which twin are you? a demonstration of induced categorical perception of identical twin faces," *British Journal of Psychology*, vol. 89, pp. 39–57, 1998.
- [14] K. R. Livingston, J. K. Andrews, and S. Harnad, "Categorical perception effects induced by category learning," *Journal of Experimental Psychology*, vol. 24, no. 3, pp. 732–753, 1998.
- [15] J. E. Pegg and J. F. Werker, "Adult and infant perception of two english phones," *Journal of the Acoustical Society of America*, vol. 102, pp. 3742–3753, 1997.
- [16] O. Rosenthal, Personal Communication, January 2008.
- [17] W. M. Jenkins, M. M. Merzenich, M. T. Ochs, T. Allard, and E. Guíñ-Robles, "Functional reorganization of primary somatosensory cortex in adult owl monkeys after behaviorally controlled tactile stimulation," *Journal of Neurophysiology*, vol. 63, no. 1, pp. 82–104, 1990.