Modeling Unsupervised Perceptual Category Learning

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

***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 (see Vallabha *et al.* in the reference section). The model treats categories as Gaussian distributions, proposing both the number and the 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 several results: acquired distinctiveness between categories and acquired similarity within categories, a faster increase in discrimination for more acoustically dissimilar vowels, and gradual unsupervised learning of category structure in simple visual stimuli.**

***Index Terms*—human learning, mixture of Gaussians, online learning, unsupervised learning.**

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 that it has feathers, it is edible, and it can fly. A long history of modeling work has investigated how categories are learned, e.g., [1]–[3].

While category learning is often facilitated by associating objects with category labels, categories can also 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 nonnative contrasts decreases [4] and sensitivity to native contrasts increases [5]. In the visual modality, Rosenthal *et al.* [6] found that subjects were sensitive to the cluster structure of the stimuli in assigning them different categories without feedback.

We consider how category structure might be learned without labels in situations posing the following additional 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 [7]. The recent online mixture estimation (OME) algorithm [7] addresses these challenges. It learns a generative model of a sequence of stimuli using a mixture of Gaussian categories, proposing both the number and parameters of the categories. OME can be seen both as an extension of competitive learning models [8] and as an online variant of expectation maximization (EM). The model is also somewhat biologically plausible since a topographical network can serve as an approximation [7].

Past work has demonstrated OME's power as a learning algorithm. In [7], the OME algorithm successfully learned the number and parameters of multidimensional vowel categories. The model used training data from the speech of Japanese- and English-speaking mothers to their children. OME attempted to learn four vowel categories from either language and was successful in most cases. However, successfully learning vowel categories does not imply the algorithm successfully captures unsupervised category learning performance of human learners. In this paper, we address this issue by comparing the OME model to several findings on both the time course and the outcome of learning from the human learning literature.

We begin by addressing a key qualitative feature of human learning, seen in both supervised and unsupervised learning situations: category learning is often marked by changes in discrimination. There is evidence that learning leads to improved discrimination across category boundaries *(acquired distinctiveness)* [5], [9], [10] and decreased discrimination within category boundaries *(acquired similarity)* [4], [11], terms from [9]. In our first simulation, OME showed these effects and thus captures two important aspects of the learning as seen in human subjects. This is consistent with past modeling work. In a related gradient ascent model formulated for one-dimensional stimuli, McMurray *et al.* [12] found both acquired distinctiveness and acquired similarity.

After replicating this basic effect, we apply OME to two other results from studies of unsupervised category learning in humans. First, we find that acoustically more distinct vowels acquire distinctiveness faster, consistent with [13]. Secondly, we apply OME to an unsupervised visual category learning task where both the time course and outcomes of learning were investigated [6]. We find that OME captures the subjects' gradual learning of the categories. Furthermore, we show OME can also model the reaction time of categorization judgements, accounting for an edge effect where a previous model failed [6].

II. The Online Mixture Estimation Model

The OME algorithm treats categories as multivariate Gaussian distributions and gradually estimates the category structure from a sequence of stimuli. The model is initialized with many (50 or more) initial *guess categories* randomly spread over the input space. After learning is complete, the

Manuscript received February 09, 2009; revised April 03, 2009. First published April 28, 2009; current version published May 29, 2009.

B. M. Lake and J. L. McClelland are with the Department of Psychology, Stanford University, Stanford, CA 94305 USA (e-mail: brenden@stanford.edu).

G. K. Vallabha was with the Department of Psychology, Stanford University, Stanford, CA 94305 USA. He is now with The MathWorks, Inc., Natick, MA 01760 USA.

Digital Object Identifier 10.1109/TAMD.2009.2021703

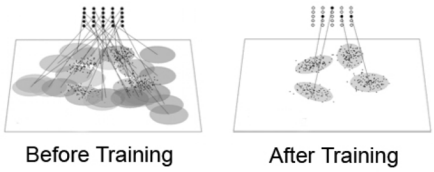

Fig. 1. OME learning stimuli in two-dimensional space. OME is initialized with many guess categories spread over input space (gray ovals), with equal mixing probabilities (grid of black circles). Learning involves presenting the stimuli (black dots) one by one and updating the categories. After learning, unneeded guess categories have mixing probabilities near zero (open circles in grid), and OME has fit the remaining guess categories to the stimulus clusters. (Reprinted with permission from [7]; copyright National Academy of Sciences, 2007.)

model proposes both the number of categories and their parameters (Fig. 1).

Learning occurs online, such that the model parameters are updated for every stimulus item D. When an item D is presented, the first step is to implicitly categorize that item. This is accomplished by calculating the posterior probability $p(r \mid D)$ for each category r with Bayes' rule, which we refer to as the *responsibility* of category r (Resp_r) for stimulus D

$$\text{Resp}_r = p(r \mid D) = \frac{p(D \mid r)\text{mix}_r}{\sum_r p(D \mid r)\text{mix}_r}. \quad (1)$$

In this calculation, $p(D \mid r)$ is the likelihood associated with the Gaussian category r. The mixing probability mix_r (also denoted as $p(r)$) denotes the probability that r contributes a random token to the stimulus sequence.

Now that the category responsibilities are calculated, the parameters of the model are updated. Each category r is parameterized by a mean M_r, a covariance matrix C_r, and the mixing probability mix_r. For each category r, M_r, and C_r are adjusted with a local update rule to better account for the stimulus D, with the update size proportional to the responsibility $p(r \mid D)$ and a learning rate. In contrast, mix_r is updated in a winner-takes-all fashion, where the winning category has the largest responsibility for D. For a precise description of the OME algorithm, see Appendix A.

The winner-takes-all update of mixing probability has been found to encourage competition amongst the categories and more effective learning. In a related gradient ascent model [12], which we discuss further in the next section, graded updates in mixing probability across the categories cause poor convergence to the correct number of categories. While the model could be using a distributed representation of the stimulus density, McMurray *et al.* found that the fit was not good [12]. If the mixing probabilities are never updated and all guess categories maintain equal weight, we have found in unreported explorations that the algorithm does not represent the stimulus density well. Alternatively, if M_r and C_r are updated only for the winning category (a fully winner-takes-all model), unreported simulations show that the model failed to learn vowel categories as successfully when compared with [7].

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

$$\text{Discrimination}(D_a, D_b) = \sqrt{\sum_r (p(r \mid D_a) - p(r \mid D_b))^2} \quad (2)$$

which is the Euclidean distance between the responsibility vectors for two stimuli D_a and D_b (the same approach was taken in [12]). 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.

III. Relationship to Past Work

Our primary aim is to investigate whether OME captures qualitative aspects of the human learning data, since our past work has only evaluated whether OME could solve the required learning problem [7]. While OME was specifically applied to learning vowel categories, more generally, OME fits a mixture of Gaussians to a set of stimuli. Other approaches to this problem exist. For instance, de Boer and Kuhl [14] used EM to learn vowel categories, which is a standard algorithm for fitting a mixture of Gaussians. Both OME and EM can solve the required learning problem and have been specifically applied to learning vowel categories. To further evaluate their plausibility as cognitive models, comparing the models to human performance is a natural next step.

A priori, there are reasons to recommend OME as a candidate for modeling the human learning process. OME learns the number of categories from the data—this does not need to be prespecified. Furthermore, OME learns online, updating after each stimulus. In contrast, EM as used in [14] repeatedly cycles through the data, and the number of categories is specified in advance. We agree with others who have argued that repeatedly cycling through a data set is not cognitively plausible, and in the tasks we consider, the number of categories is not provided to the learner in advance [14].

There are other approaches, such as stochastic gradient ascent, that fulfill these desirable properties. In fact, OME is related to maximum likelihood estimation by stochastic gradient ascent. Working independently from our group, McMurray *et al.* [12] formulated a model for one-dimensional stimuli, deriving gradient ascent updates for the category means, variances, and mixing probabilities. After modifying the model for winner-takes-all mixing probability updates, their model and OME are closely related. For one-dimensional stimuli, the mean and variance updates are roughly the same, except that with gradient ascent, the current variance of a category affects the magnitude of the updates (see Appendix B for details). While the authors derived the model for only one-dimensional stimuli, it could presumably be derived in the general multidimensional case. Further work is needed to see if these two models make different quantitative predictions. Currently, we see them as sharing the same critical properties: unsupervised learning, online updating,

TABLE I
TRAINING FREQUENCIES USED IN EXPERIMENT 1

Token #	1	2	3	4	5	6	7	8
Bimodal	4	16	8	4	4	8	16	4
Unimodal	4	4	8	16	16	8	4	4

estimating the number of categories from data, and enforcing competition amongst the categories.

There are other models that also fulfill these desired properties, such as the rational model of categorization (RMC) [2] and SUSTAIN [3]. While this paper does not attempt to discriminate between these alternative approaches, we discuss this as a future direction in Section VII.

IV. EXPERIMENT 1: LEARNING ONE VERSUS TWO CATEGORIES

Changes in discrimination can be used to investigate the category learning process. In this first experiment, we looked at two central effects: acquired distinctiveness (improving discrimination across category boundaries [5], [9], [10]) and acquired similarity (declining discrimination within category boundaries [4], [11]). In McMurray *et al.*'s related model [12], both effects were found when learning the bimodal voice onset time distribution in English. Experiment 1 serves to replicate this result with OME, specifically addressing the task and findings of an experiment by Maye *et al.* [15] investigating learning from speech sounds presented to infants.

In the infant study [15], 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 [16]. The continuum in the experiment was drawn from either a unimodal or a bimodal distribution for each infant (Table I), and the authors hypothesized that the infants would form either a one-category or a two-category representation, respectively. From this hypothesis and from the principles of acquired distinctiveness and acquired similarity, they predicted that infants exposed to the bimodal distribution would show better discrimination on the contrast than those exposed to the unimodal distribution.

Post-sensitization testing confirmed their prediction. Infants in both sensitization groups received two types of test trials: alternating (a string of eight stimuli alternating between the endpoint tokens 1 and 8 in Table I) and nonalternating (a string of eight identical tokens). Infants sensitized to the bimodal distribution showed a significant difference in looking time between the test trial types, while infants sensitized to the unimodal distribution did not. This indicates that infants in the bimodal condition were sensitive to the contrast, and those in the unimodal condition were less sensitive to it.

For the model, we trained OME on either the unimodal or bimodal distribution used in the study (Table I). If OME shows acquired distinctiveness, exposure to the bimodal distribution will increase discrimination of the speech contrast over the course of learning. If OME also shows acquired similarity, exposure to the unimodal distribution will decrease discrimination over the course of learning.

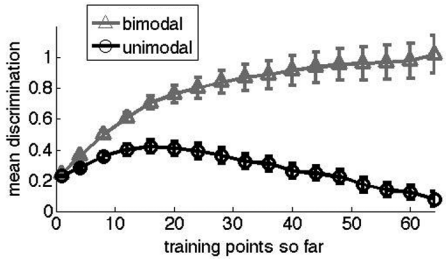

Fig. 2. Time course of [da]-[ta] discrimination for the two distributions averaged over runs. A bimodal stimulus distribution causes an increase in discrimination over time (gray line), while a unimodal distribution causes a decrease in discrimination over time (black line). The error bars are standard error.

A. Model

Both the infants and model were presented with 64 stimulus tokens. The stimuli presented to the model were simply the values one through eight, presented in random order. To simulate multiple subjects, 24 models with different random initializations were run for each distribution. The parameters used to initialize the model are listed in Appendix C and Table III. The category centers and standard deviations were initialized randomly so that the structure of the actual data was not anticipated. Using the discrimination metric from (2), we measured discrimination of the [da]-[ta] continuum endpoints (one versus eight) as training progressed.

B. Results

OME was largely successful at finding the correct structure, although six models trained on the bimodal distribution formed only a single category (whether a subset of human participants likewise failed to discover two categories in this condition is not known). After learning, models trained on the bimodal distribution showed significantly better discrimination than those trained on the unimodal distribution $(t(46) = 7.12, p < .001)$. When compared before and after training, models exposed to the bimodal distribution increased in discrimination $(t(23) = 6.4, p < .001)$, and those exposed to the unimodal distribution decreased in discrimination $(t(23) = 3.27, p < .01)$. The time course of discrimination is plotted in Fig. 2.

The model accounts for two aspects of the infant data. First, when comparing the two distributions, both infants and the model exposed to the bimodal stimuli showed better discrimination. Secondly, the infants who were familiarized to the unimodal distribution did not significantly discriminate the endpoints in [15], although a past study has shown infants of this age can make this discrimination [16]. Thus, the unimodal distribution likely caused a reduction in discrimination. The OME model shows this reduction. By accounting for the data, OME shows how the learning process could occur—through small, online updates to the category structure as the infant receives speech tokens.

More generally, OME demonstrates acquired distinctiveness and similarity effects. This was originally shown in the McMurray *et al.* [12] model, and thus OME replicates the effects. Both effects have empirical evidence from supervised and unsupervised category learning in various modalities [4], [5], [9]–[11] and follow naturally from the modeling framework.

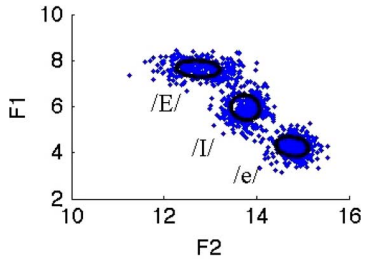

Fig. 3. Vowel categories estimated from Sabourin [13] and converted to the Bark scale. The 1000 gray points are a random draw, and the black circles are the categories OME found, plotted 1 standard deviation along each principal axis.

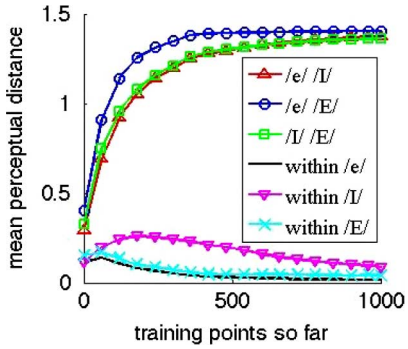

Fig. 4. This figure plots perceptual distance across training, averaged over ten runs. The largest standard error, if shown, would be 0.036. In the legend, the first three lines illustrate increasing discrimination between vowels and the last three lines show decreasing discrimination within vowels. The more distinct /e/ versus /E/ contrast differentiates first.

Acquired distinctiveness is demonstrated by the gray line in Fig. 2. As the model forms two categories, discrimination of the contrast increases. This is because some guess categories may overlap with both sides of the contrast initially until two nonoverlapping categories dominate. Acquired similarity is demonstrated by the black line. A single category provides no discrimination with our metric, so the initial random configuration supports better discrimination. Interestingly, the model predicts an initial increase in within-category discrimination before the eventual decrease.

V. Experiment 2: More Widely Separated Categories Acquire Distinctiveness Faster

We next consider the effect of the spacing of stimulus clusters on unsupervised category learning. A study by Sabourin *et al.* [13] addressed this and found better discrimination of more widely separated clusters compared to clusters placed closer together. Here we show that OME also shows this same effect.

We model Sabourin *et al.*'s [13] study that tested eight-month-old English monolinguals on their ability to discriminate vowels /e/ versus /I/ and /e/ versus /E/ (Fig. 3). The vowels were presented in a /tVb/ frame, and infants were initially habituated to the stimulus /teb/. Each infant was then tested on one of the possible discriminations. Test trails consisted of Same trials (new instances of /teb/) or Different trials (instances of either /tIb/ or /tEb/). The infants showed no behavioral evidence of discriminating the closer vowels (/e/ versus /I/) and only weak support for /e/ versus /E/. Further investigation using event-related potentials indicate some sensitivity to both contrasts, with greater sensitivity to /e/ versus /E/ than /e/ versus /I/.

We present a simulation of learning in OME that can account for these findings. In this experiment, we trained the model on points drawn from these three vowel categories (Fig. 3) and tracked discrimination between the categories. The stimuli were treated as points in a two-dimensional auditory space, corresponding to F1 and F2 space. We would expect discrimination to increase faster between /e/ versus /E/ than /e/ versus /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 ten different times with different draws from the vowel stimuli (Fig. 3). See Appendix D for details on generating the vowel stimuli and Appendix C (Table III) for the model parameters.

A. Results

To calculate discrimination between two categories rather than two stimuli as in (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, 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 three vowel categories in all ten runs. As in Experiment 1, the model showed acquired distinctiveness and similarity (Fig. 4). Thus these basic effects are also found when learning multidimensional categories. We also found that the model's discrimination of vowels is affected by acoustic distance (Fig. 4). In particular, the perceptual distance between vowels /e/ versus /E/ grew significantly faster than /e/ versus /I/, accounting for the result in [13].[1] In OME, further apart categories differentiate faster, consistent with the infant data.

VI. Experiment 3: Visual Category Learning

In the past two experiments, discrimination changed with training, consistent with empirical findings. However, OME has not yet been compared to experimental data where performance was evaluated at multiple points throughout learning. In the next experiment, we modeled data from an unsupervised visual categorization task that provides data from multiple time points.

[1]To check the significance of the faster growth, we fit a damped exponential curve to each distance trajectory separately for the ten 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/ versus /E/ differentiates significantly faster, $t(9) = 7.39, p < .001$.

While previous modeling with OME has been auditory, OME's principles can also be applied to visual category learning: unsupervised learning and online updating, without specification of the number of categories.

OME was applied to Rosenthal *et al.*'s [6] 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 [three peaks, four peaks, or a uniform distribution with no peaks; these distributions are illustrated as the black curves in Fig. 5(a)]. Learning was divided into four sessions, and learning progress was evaluated at the end of each session. 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 near stripe widths that appeared most frequently.

To account for this implicit learning, 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 a wrong prediction using their model network's 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 Rosenthal *et al.* model showed the opposite. In this experiment, we show that the OME algorithm can account for this aspect of the data.

This unsupervised learning experiment is a much stronger test of the adequacy of the OME approach, due to (a) a comparison of learning three versus four categories, (b) observation of learning performance session by session, and the additional (c) availability of RT data. Each of these aspects provides a novel test for OME compared with the previous two experiments.

A. Model

OME was trained for 4096 trials, as were subjects. There were 20 replications for each condition (three-peak, four-peak, and uniform) with different random stimulus sequences and starting guess categories. Rosenthal *et al.* [6] allowed the stripe width to vary from one to 512 pixels, divided into 36 sample bins. The model stimuli were integers from one to 36 drawn from the same distributions, then perturbed by Gaussian noise with standard deviation 2.3 before training OME. See Appendix C (Table III) for the model parameters. Subjects were shown 256 trials that preceded the main experiment to acquaint them with the task and distribution. Subjects pressed the same key in these trials. The model was given an analogous 256 trials before data was recorded.

B. Results

Both the OME model and human subjects organized their category structures based on the distributional properties of the stimuli. Learning was separated into four sessions in the experiment, consisting of 1024 trials each. OME's performance was evaluated in the last session and throughout the sessions. Fig. 5 displays the behavior and model side by side.

Evaluated at the end of learning, OME and subjects learned similar category structures. During the fourth session (last quarter of training), the subjects and model showed a clear tendency to organize category centers near peak frequencies and category boundaries between peak frequencies. As seen in Fig. 5(a), both subjects and models exposed to the three-peak distribution learned category centers around those peaks, and likewise for the four-peak distribution. Boundaries were learned between peaks. For subjects and models exposed to the uniform distribution, the learned structure was more arbitrary. Note that the subjects tended to place extra categories at the extremes of the stimulus range while OME did not. See Appendix E for details on calculating the centers and boundaries. The number of classes found varied across subjects (SD > 2 for the fourth session in each condition) and also across different runs of the model (SD> 0.7), though there was more variability across subjects than across runs of the model.

RT was also evaluated at the end of learning. Previous simulations with OME have only evaluated the category structure learned and discrimination, so RT provides a novel test for OME. OME's performance in strikingly similar to the behavior. To simulate RT in the model for a stimulus, we took the largest guess category responsibility [$p(r \mid D)$ calculated in (1)] minus the second largest (best minus next), which is a measure of confidence and is thus inversely related to RT. The confidence score was fit to the subject RT data with linear regression. If one guess category was clearly the most responsible for a stimulus, the responsibility difference is large (confidence is large) and RT is small. If two guess categories were similarly responsible, the responsibility difference is small and the RT is large. Fig. 5(b) compares RT in the subjects and the model for the fourth session, finding a clear tendency for longer RT near the midpoint between the category centers.

Interestingly, Rosenthal *et al.* [6] found that stimuli near the edges of the range (very narrow or wide lines) were categorized faster (particularly for the four-peak and uniform cases). Rosenthal *et al.*'s neural network model of the task makes the opposite prediction using settling time as a measure of RT. The model is a layer of fully interconnected units utilizing Hebbian learning, where a stimulus excites consecutive neurons in up to 1/3 of the network. When presented with a peaked stimulus distribution, the network forms an attractor for each peak. When stimulus settling time is interpreted as RT, extreme stimuli (like boundary stimuli) take considerable time to settle into an attractor. This leads the Rosenthal *et al.* model to make the wrong prediction [Fig. 5(b)].

OME makes the correct prediction regarding this effect [Fig. 5(b)]. The guess category situated closest to the edge of the stimulus space would have nearly all the responsibility for extreme stimuli near that edge. This category will have little competition for these extreme stimuli, producing a small "second largest category responsibility" and consequently a low simulated RT. Because of this, OME accounts for the edge effect.

Rosenthal *et al.* [6] also investigated the time course of categorization. In Fig. 5(c), the histogram displayed in Fig. 5(a) was calculated for each of the four sessions and averaged across multiple peaks (edge stimuli were removed for consistency with

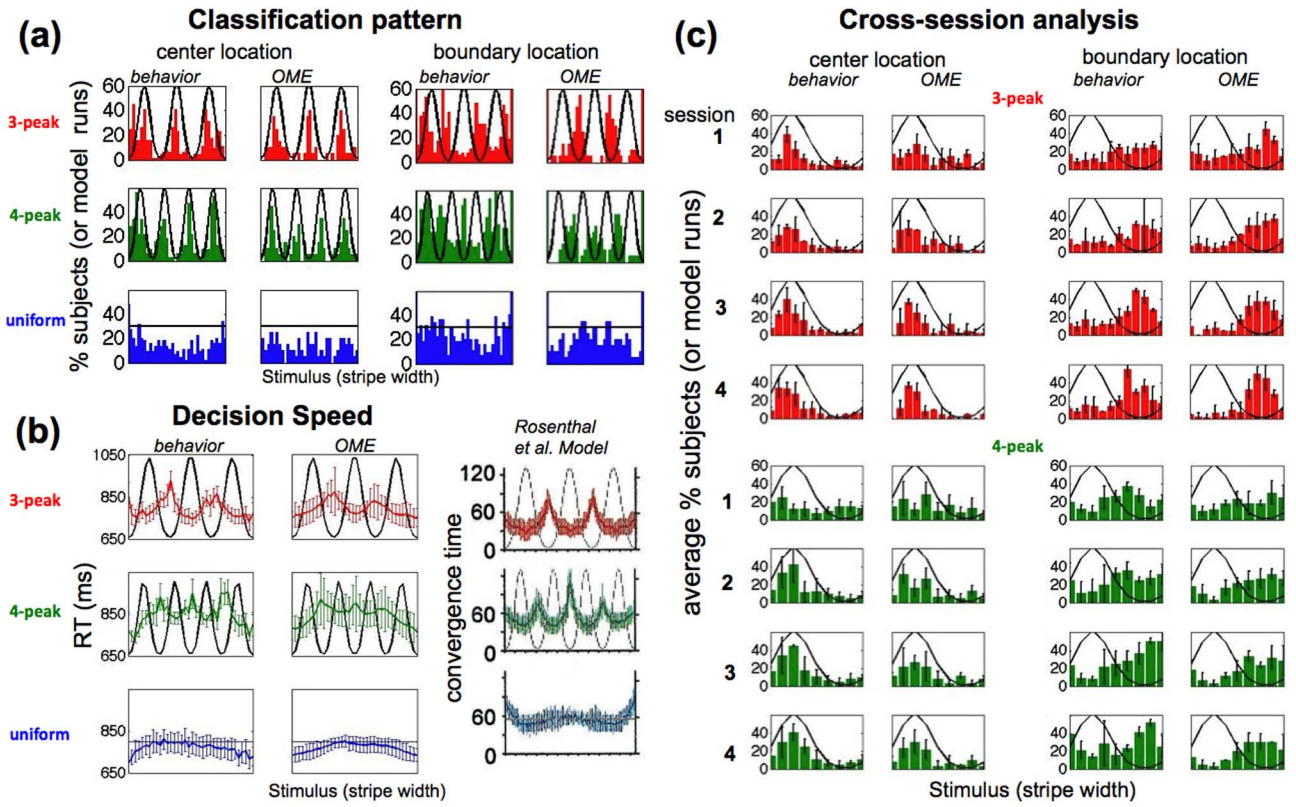

Fig. 5. All figures labeled "behavior" use subject data from an experiment in [6]. (The image under "Rosenthal model" is reprinted with permission from [6], copyright National Academy of Sciences, 2001.) Note that some bars in this figure extend beyond 60%. (a) For the fourth session, this figure shows 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) having 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 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 model, which increases at the edges where the behavior and OME do not. The error bars are standard error for the behavior and standard deviation for the models for consistency with [6]. OME RT was fit with linear regression with parameters (three-peak, four-peak, uniform): RT $= 1071 - 336X$, RT $= 1165 - 421X$, and RT $= 904 - 192X$. (c) 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. Error bars are standard deviation.

[6]). This plot illustrates how learning evolved across sessions. As with the subjects, the model's centers and boundaries were increasingly influenced by the stimulus distribution as the sessions progress. A particular feature of the behavioral data is that the three-peak distribution is learned faster, evident in both the centers and boundaries. OME also shows this effect, which is related to the noise parameter in the model. Since the categories are better separated in the three-peak distribution, high noise in the stimuli hinders learning more dramatically in the four-peak distribution. Although the model shows this effect, the fit to the details of the time course data is not perfect; the subjects reach near asymptotic performance levels relatively quickly, while the model improves more gradually over sessions.

The behavior of OME is not independent of its free parameters, although OME accounts for the basic effects in this simulation across a substantial range. A particularly important parameter is the initial width standard deviation (s.d.) of the guess categories. If the width is too large, such that all 100 initial categories begin with an s.d. of five or larger (where the total stimulus range was 35), OME incorrectly merges categories. In the presented results, the initial s.d. of each guess category was picked randomly from 0.97 to 48.5 (Table III). Since the initial categories can vary widely in width, and many are too large to be viable, both the number and initial placement of viable categories are random. With these parameters, the model often, but not always, recovers the correct category structure. We find this best approximates the aggregate behavioral results.

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' RT, it seems natural that OME would be more certain about a categorization query for peak and edge stimuli.

VII. General Discussion

Categorization is essential to perception, and much of perceptual category learning is unsupervised. How can category structure be learned from just a sequence of stimuli? The OME algorithm [7] 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 [7] does not show the algorithm is a model of the process as it occurs in human learners.

From this work, there are several results to recommend OME as a model of human category learning. In both Experiments 1 and 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 and follow naturally from the modeling framework. To further investigate how discrimination develops over time in Experiment 2, we found that OME acquired distinctiveness faster for more acoustically dissimilar vowels, consistent with infant data [13]. Experiment 3 provided the strongest test of OME, due to a comparison of three versus four categories, observation of learning performance session by session, and the availability of RT data. OME provided an account for this rich array of data, including an edge effect in RT that a previous model failed to show. OME's success demonstrates that the same principles governing auditory category learning in the model can be applied to visual category learning.

A. Advantages of OME

More generally, OME provides an elegant solution to the problems of 1) scalability, 2) revisability, and 3) cross-modal fusion in category learning. Regarding scalability, OME's computational complexity is $O(nRd^2)$ with n training examples, R initial guess categories, and d stimulus dimensions. Speed is independent of the number of actual categories and is strictly linear in the number of training examples, using each one in turn to provide a gradually evolving estimate of the category structure in the training data. Speed can be further improved by deleting guess categories with low probability during learning. Furthermore, with high-dimensional data, a restriction to guess categories with diagonal covariance matrices reduces the complexity to $O(nRd)$.

Although not shown in the simulations we have presented here, the model is able to revise its solution if presented with a changing profile of category structure over the course of learning. 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, an unused guess category can be recruited, provided these have not been eliminated during the learning process.

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.

OME can also be approximated by a nonparametric algorithm "topographic OME" (TOME, from [7]) that moves closer to a possible neurobiological interpretation and removes the strong Gaussian constraint on categories. Instead of a Gaussian, a category is represented nonparametrically by dividing the input space into regions and estimating the proportion of inputs in each region. This representation scheme has a "neural network" interpretation: the proportions can be encoded as connection weights between neuron-like units standing for small regions of the input space and units standing for category representations. We refer interested readers to [7] for the details of this algorithm.

B. Future Directions

Further research is needed to understand how people estimate the number of stimulus clusters during unsupervised category learning. OME takes a unique approach to solving this problem, starting with an initial overabundance of guess categories. With the exception of the related gradient ascent model [12], alternative models such as the RMC [2] and SUSTAIN [3] often start with one guess category and add additional ones as necessary. Future work will investigate whether this difference leads to different behavioral predictions.

What if the category learning problem is semisupervised rather than unsupervised, where both labeled and unlabeled category examples are presented as stimuli? This more accurately describes many types of naturalistic learning problems such as learning object categories. A further complication of labeled feedback is that Gaussian categories are no longer a reasonable simplification. For example, the visual category "boat" includes both sailboats and motorboats, which is certainly bimodal with regard to sail-related features. If labels are introduced, very different types of categories can be learned. Can OME be extended to semisupervised learning? By representing a category label as an additional binary feature of each data point [2], labeled categories can be represented across a collection of Gaussians in OME. Future work will explore this approach.

Another interesting issue is how repeated stimulation affects sensitivity. In [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 show increased sensitivity within regions densely populated by presented stimuli while others show decreased sensitivity within such regions. We are examining whether modifications to OME could produce the opposite behavior, potentially providing insight into this question.

APPENDIX

A. Operation of OME

OME is initialized with R Gaussian distributions, each parameterized by a mean M_r, covariance matrix C_r, and mixing probability mix_r. The mixing probabilities mix_r are initialized to be $1/R$, and each M_r and C_r is initialized randomly within a certain range.

On each trial, the algorithm goes through six steps, summarized as follows.

1) Get the input stimulus D.
2) Calculate the likelihood of D for each category r multiplied by the prior probability mix_r.
3) Calculate the responsibility for each category r.
4) Update the parameters for each category r.
5) Update the mixing probability for winning category $\hat{r}$.
6) Ensure mixing probabilities sum to 1.

TABLE II
OME VERSUS SGA

	OME	SGA
ΔM_r	$\eta Resp_r(D - M_r)$	$\eta Resp_r(D - M_r)/\sigma_r^2$
$\Delta\sigma_r^2$	$\eta Resp_r((D - M_r)^2 - \sigma_r^2)$	$\eta Resp_r((D - M_r)^2 - \sigma_r^2)(2/\sigma_r^2$ $+\eta Resp_r((D - M_r)^2 - \sigma_r^2)/\sigma_r^6)$

TABLE III
OME INITIALIZATION

	R	M_r	σ_r	η	η_{wta}	$delete$
Exp. 1	50	-1 to 10	2 to 4	.05	.05	.001
Exp. 2 (F1)	50	3.9 to 8	.5 to 1	.01	.01	.001
Exp. 2 (F2)		12.4 to 15	.5 to 1.5			
Exp. 3	100	1 to 36	$\rho/2$ to 25ρ	.005	.0025	.0001

Below, Resp_r is the responsibility of category r for data point D.

1: Get a data point D
2: For $\text{r} = 1 \ldots R, p_r \leftarrow \text{mix}_r \cdot P(D \,|\, r; M_r, C_r)$
3: For $\text{r} = 1 \ldots R, \text{Resp}_r \leftarrow p_r / \sum_j p_j$
4: For $\text{r} = 1 \ldots R$

$$M_r \leftarrow M_r + \eta \cdot \text{Resp}_r \cdot (D - M_r)$$
$$C_r \leftarrow C_r + \eta \cdot \text{Resp}_r \cdot [(D - M_r) \cdot (D - M_r)^T - C_r]$$

5: $\hat{r} \leftarrow \arg\max\{\text{Resp}_r\}; \text{mix}_{\hat{r}} \leftarrow \text{mix}_{\hat{r}} + \eta_{wta}$
6: For $\text{r} = 1 \ldots R, \text{mix}_r \leftarrow \text{mix}_r / \sum_j \text{mix}_j$

B. Comparing OME to Stochastic Gradient Ascent

As described in Section III, OME is related to a model by McMurray *et al.* [12] for one-dimensional stimuli derived by stochastic gradient ascent (SGA). When compared with OME in one dimension, the operation of the algorithms is largely the same. The updates of the category means and variances are related, but not identical [Step 4) above]. Table II specifies the difference in parameter updates. ΔM_r denotes the update to the mean of category r and $\Delta\sigma_r^2$ the update to the one-dimensional variance. Notice that in SGA, as the variance of a category r increases, the magnitude of the parameter updates decreases.

C. Initialization of Model Parameters

The parameters for each simulation are listed in Table III. The initial guess category means M_r and standard deviations σ_r are randomly drawn from a uniform distribution in the range indicated. The other parameters are as follows: R is the number of guess categories, η is the learning rate for means and covariance matrices, η_{wta} is the learning rate for mixing probability, and guess categories that fall below a mixing probability of *delete* are removed for efficiency. In Experiment 2, the categories were initialized with no off-diagonal terms in the covariance matrix as indicated for the dimensions F1 and F2. In the Experiment 3 row of the table, $\rho = 1.94$ and is the standard deviation of a peak in the three-peak frequency distribution (for comparison, the s.d. of a four-peak is 1.46).

D. Vowel Stimuli for Experiment 2

We drew vowel data for /e/, /E/, and /I/ from Gaussian distributions to be learned by OME. The authors [13] 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 off-diagonal terms in the covariance matrix. With these estimated parameters, we drew a total of 1000 points with equal probability from each vowel. The points were then converted to the Bark scale (1 to 24, corresponding to the first 24 critical bands of hearing). We ran ten simulations with different draws for the vowel points and initial categories for OME.

E. Calculating Category Centers and Boundaries in Experiment 3

Fig. 5(a)–(c) relies on the "centers" and "boundaries" of the categories learned by subjects (or by OME) during the experiment. For the subjects, these clearly must be inferred from the category responses. While we could directly find the centers and boundaries in OME by looking at the learned category parameters, the procedure used for the subjects was followed for closer comparison.

The procedure for calculation of centers and boundaries followed [6]. First, the authors defined a sorting coherence function, for a particular subject, $\chi(b, r)$ as the fraction of presentations a stimulus in bin b was classified as class r. To simulate this measure in OME, we defined a classification for a particular stimulus as the guess category with the highest responsibility (1). Then, calculating the sorting coherence function for a session is straightforward. Secondly, the center and boundaries must be calculated for each category r chosen for classification in a particular session. The center was defined as $\sum_{b=\text{left boundary}}^{\text{right boundary}} b \cdot \chi(b, r) / \sum_{b=\text{left boundary}}^{\text{right boundary}} \chi(b, r)$. The location of the left (right) boundary was the last bin in which $\chi(b, r) \leq 0.5$, when starting at the leftmost (rightmost) bin and moving towards the right (left). The edge bins were the boundaries for the edge categories [18].

ACKNOWLEDGMENT

The authors gratefully thank O. Rosenthal for her comments and for providing data used in Fig. 5. They also thank three anonymous reviewers for their helpful suggestions.

REFERENCES

[1] D. L. Medin and M. M. Schaffer, "Context theory of classification learning," *Psychol. Rev.*, vol. 85, no. 3, pp. 207–238, 1978.
[2] J. R. Anderson, "The adaptive nature of human categorization," *Psychol. Rev.*, vol. 98, pp. 409–429, 1991.
[3] B. C. Love, D. L. Medin, and T. Gureckis, "Sustain: A network model of category learning," *Psychol. Rev.*, vol. 111, no. 2, pp. 309–332, 2004.
[4] J. F. Werker and R. C. Tees, "Cross-language speech perception: Evidence for perceptual reorganization during the first year of life," *Infant Behav. Develop.*, vol. 7, pp. 49–63, 1984.
[5] 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," *Develop. Sci.*, vol. 9, no. 2, pp. F13–F21, 2006.
[6] O. Rosenthal, S. Fusi, and S. Hochstein, "Forming classes by stimulus frequency: Behavior and theory," *Proc. Nat. Acad. Sci.*, vol. 98, pp. 4265–4270, 2001.
[7] G. K. Vallabha, J. L. McClelland, F. Pons, J. F. Werker, and S. Amano, "Unsupervised learning of vowel categories from infant-directed speech," *Proc. Nat. Acad. Sci.*, vol. 104, no. 33, pp. 13 273–13 278, 2007.
[8] D. E. Rumelhart and D. Zipser, "Feature discovery by competitive learning," *Cogn. Sci.*, vol. 9, no. 1, pp. 75–112, 1985.

[9] R. K. Goldstone, "Influences of categorization on perceptual discrimination," *J. Exper. Psychol., Gen.*, vol. 123, no. 2, pp. 178–200, 1994.
[10] S. V. Stevenage, "Which twin are you? A demonstration of induced categorical perception of identical twin faces," *Br. J. Psychology.*, vol. 89, pp. 39–57, 1998.
[11] K. R. Livingston, J. K. Andrews, and S. Harnad, "Categorical perception effects induced by category learning," *J. Exper. Psychol.*, vol. 24, no. 3, pp. 732–753, 1998.
[12] B. McMurray, R. N. Aslin, and J. Toscano, "Statistical learning of phonetic categories: Insights from a computational approach," *Develop. Sci.*, 2009.
[13] L. Sabourin, J. F. Werker, L. Bosch, and N. Sebastián-Gallés, "Perceiving vowels in a tight vowel space: Evidence from monolingual infants," *Develop. Sci.*, submitted for publication.
[14] B. de Boer and P. K. Kuhl, "Investigating the role of infant-directed speech with a computer model," *Acoust. Res. Lett. Online*, vol. 4, pp. 129–134, 2003.
[15] J. Maye, J. F. Werker, and L. Gerken, "Infant sensitivity to distributional information can affect phonetic discrimination," *Cognition*, vol. 82, pp. B101–B111, 2002.
[16] J. E. Pegg and J. F. Werker, "Adult and infant perception of two english phones," *J. Acoust. Soc. Amer.*, vol. 102, pp. 3742–3753, 1997.
[17] W. M. Jenkins, M. M. Merzenich, M. T. Ochs, T. Allard, and E. Guíc-Robles, "Functional reorganization of primary somatosensory cortex in adult owl monkeys after behaviorally controlled tactile stimulation," *J. Neurophysiol.*, vol. 63, no. 1, pp. 82–104, 1990.
[18] O. Rosenthal, personal communication Jan. 2008.

Brenden M. Lake will receive the B.S. and M.S. degrees in symbolic systems, specializing in computational and statistical approaches to learning and inference, from Stanford University, Stanford, CA, in 2009. Beginning Fall 2009, he will start working towards the Ph.D. degree at the Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology (MIT), Cambridge.

Gautam K. Vallabha received the B.S. degree in electrical and computer engineering from Carnegie Mellon University, Pittsburgh, PA, in 1995 and the Ph.D. degree in complex systems and brain sciences from Florida Atlantic University, Boca Raton, in 2003.

From 2003 to 2007, he was a Postdoctoral Researcher at the Center for the Neural Basis of Cognition, Carnegie Mellon University and at the Department of Psychology, Stanford University, specializing in neural network models of speech perception and language learning. Currently, he is with The MathWorks, Natick, MA.

James L. McClelland received the Ph.D. degree in cognitive psychology from the University of Pennsylvania in 1975.

After his Ph.D., he joined the faculty at University of California at San Diego (UCSD), where he, D. E. Rumelhart, and others produced the two-volume work "Parallel Distributed Processing" in which artificial neural networks employing learned distributed representations were presented as a framework for modeling the human cognition as an emergent consequence of neural activity. He (with D. E. Rumelhart) received the IEEE Neural Networks Pioneer award for this work. He moved to Carnegie Mellon University, Pittsburgh, PA, in 1984, where be became the founding Co-Director of the Center for the Neural Basis of Cognition, and then moved to Stanford University, Stanford, CA, in 2006, where he is currently Professor of Psychology. He is also Director of the Center for Mind, Brain, and Computation. He continues to explore distributed connectionist models of human cognitive processes and human learning and the relationship between these models and other computational frameworks.