A Rational Analysis of the Acquisition of Multisensory Representations

Ilker Yildirim, Robert A. Jacobs

Department of Brain and Cognitive Sciences, University of Rochester

Abstract

How do people learn multisensory, or amodal, representations, and what consequences do these representations have for perceptual performance? We address this question by performing a rational analysis of the problem of learning multisensory representations. This analysis makes use of a Bayesian nonparametric model that acquires latent multisensory features that optimally explain the unisensory features arising in individual sensory modalities. The model qualitatively accounts for several important aspects of multisensory perception: (a) it integrates information from multiple sensory sources in such a way that it leads to superior performances in, for example, categorization tasks; (b) its performances suggest that multisensory training leads to better learning than unisensory training, even when testing is conducted in unisensory conditions; (c) its multisensory representations are modality invariant; and (d) it predicts ''missing'' sensory representations in modalities when the input to those modalities is absent. Our rational analysis indicates that all of these aspects emerge as part of the optimal solution to the problem of learning to represent complex multisensory environments.

Keywords: Multisensory perception; Learning; Bayesian modeling; Rational analysis

1. Introduction

Much of the history of perceptual science can be characterized as a ''sense by sense'' approach in which each sensory modality is studied in isolation. For example, visual scientists study behavioral, cognitive, and neural aspects of visual perception, whereas auditory scientists study behavioral, cognitive, and neural aspects of auditory perception. However, it is an undeniable, but often overlooked, fact that perception is fundamentally multisensory. We learn about our environments by seeing, hearing, touching, tasting, and smelling these

Correspondence should be sent to Robert A. Jacobs, Department of Brain and Cognitive Sciences, University of Rochester, Rochester, NY 14627. E-mail: robbie@bcs.rochester.edu

environments. Moreover, evolution has designed our brains so that our senses work in concert. Objects and events can be detected rapidly, identified correctly, and responded to appropriately because our brains use information derived from different sensory channels cooperatively. This fundamental property of human perception is a key reason that human intelligence is so robust.

To date, most computational work on multisensory perception focuses on the problem of sensory integration: how to combine information from two or more sensory modalities to maximize performance on a task (e.g., Abidi & Gonzalez, 1992; Alais & Burr, 2004; Battaglia, Jacobs, & Aslin, 2003; Clark & Yuille, 1990; Gepshtein & Banks, 2003). Ernst and Banks (2002), for example, examined how people estimate the height of an object based on visual and haptic inputs. They compared people's judgments with the predictions of a statistically optimal rule based on maximum likelihood estimation theory (the Fuzzy Logical Model of Perception of Massaro, 1998, provides a similar framework for sensory integration). According to this rule, an observer first forms two estimates of object height, one based on the visual input and the other based on the haptic input. Next, the observer forms an estimate of object height based on both sets of inputs by taking a weighted average of the estimates based on the individual inputs.

Although sensory integration is an important aspect of multisensory perception, it is only one of many aspects of multisensory perception. Here, we take a broader approach to the study of multisensory perception. We are interested in providing a single, unified account of a wide variety of multisensory phenomena. We are not interested in providing an account solely of sensory integration, or of providing multiple accounts, one for each phenomenon that needs to be explained. Consequently, we emphasize the need to understand multisensory representations that can underlie many types of multisensory behaviors.

Specifically, we focus on the problem of learning multisensory, or amodal, representations from unisensory data. This focus distinguishes our work from both the maximum likelihood approach to sensory integration described above and from the Fuzzy Logical Model of Perception, neither of which is concerned with the acquisition of complex multisensory representations. The representations we seek are task independent. That is, they are not acquired to facilitate performance on a particular task. Instead, they are acquired to facilitate performances on many different tasks. When performing a particular task, the acquired representations can serve as key components of a larger system that also includes components that are task specific. Our emphasis on task-independent multisensory representations is not intended to deny the existence of task-specific representations. To the contrary, we think that both task-independent and task-specific representations will play important roles in comprehensive theories of perception. Rather, our emphasis on task-independent representations allows us to focus on general theoretical principles with broad explanatory power.

The computational model that we propose uses a probabilistic framework, specifically a Bayesian framework, for studying the acquisition of multisensory representations. Bayesian modeling has become increasingly important in the field of cognitive science (e.g., Anderson, 1990; Barlow, 1961; Chater & Oaksford, 2008; Geisler, 2004; Green & Swets, 1966; Griffiths, Kemp, & Tenenbaum, 2008; Kahneman, Slovic, & Tversky, 1982; Knill & Richards, 1996; Marr, 1982; Oaksford & Chater, 1999; Todorov, 2004). A common observation

of cognitive scientists is that we live in an uncertain world, and rational behavior depends on the ability to process information effectively despite ambiguity or uncertainty. Cognitive scientists, therefore, need methods for characterizing information and the uncertainty in that information. Fortunately, such methods are available: Probability theory provides a calculus for representing and manipulating uncertain information. To us, an advantage of Bayesian models relative to many other types of models is that they are probabilistic.

Probability theory does not provide just any calculus for representing and manipulating uncertain information; it provides an optimal calculus (Cox, 1961). Consequently, an advantage of Bayesian modeling is that it gives cognitive scientists a tool for defining rationality. Via Bayes' rule, Bayesian models optimally combine information based on prior beliefs with information based on observations or data. Via Bayesian decision theory, Bayesian models can use these combinations to choose actions that maximize task performance. Due to these optimality properties, Bayesian models perform a task as well as the task can be performed (given the assumptions built into the model), meaning that the performance of a Bayesian model on a task defines rational behavior for that task (again, based on the model's assumptions).

Our model is motivated by the need for a *computational theory* (Marr, 1982) or *rational analysis* (Anderson, 1990) of the learning problem facing humans in complex multisensory environments. This approach emphasizes that important properties of cognition can be understood by characterizing the computational demands of a natural environment, and theorizing that evolution has shaped the brain so that it efficiently meets these demands. By studying multisensory behavior from this perspective, one can examine the properties of multisensory perception that emerge as part of the optimal solution to the problem of learning to perceive and represent multisensory environments. One can then compare human performances in multisensory settings to the properties predicted by the optimal model, and explain *why* these properties exist in human behavior. Cognitive scientists are increasingly using optimal models to study human cognition. Analyses based on optimal performance are referred to as rational analyses, ideal observer analyses, or ideal actor analyses in the literatures on cognition, perception, and motor control, respectively (e.g., Anderson, 1990; Geisler, 2004; Todorov, 2004).

This article is organized as follows. Section 2 describes four hypotheses about multisensory perception that have appeared in the scientific literature. We highlight these hypotheses because they play important roles in our computer simulations. Sections 3 and 4 provide preliminary remarks regarding the proposed model and a detailed description of the model, respectively. Sections 5–7 provide simulation results on three data sets: a synthetic data set, a real-world visual-auditory data set, and a real-world visual-haptic data set. The final section provides a general discussion and concluding remarks.

2. Four hypotheses about multisensory perception

Cognitive neuroscientists have recently studied at least four important hypotheses about multisensory perception. First, researchers have conjectured that multisensory representations

are essential because sensory integration provides significant statistical advantages (e.g., Abidi & Gonzalez, 1992; Alais & Burr, 2004; Battaglia et al., 2003; Clark & Yuille, 1990; Ernst & Banks, 2002; Gepshtein & Banks, 2003).

For example, sensory integration can ameliorate the effects of noise contained in representations based on single modalities. Multisensory representations are, therefore, able to convey more accurate and reliable information than the unisensory representations from which they are derived. Consider an observer that sees and touches a surface slanted in depth. Suppose that the observer's slant estimates based on the visual cue and on the haptic cue are each corrupted by sensory noise with some variance. It is easily shown that the statistically optimal estimate of surface slant obtained by combining information from both cues has a lower variance, and is thus more reliable, than estimates based on either cue alone.

Evidence that the brain is able to combine sensory information in such a manner has been obtained by several researchers. Ernst and Banks (2002) found that people's estimates of object height based on both visual and haptic information were more reliable (i.e., had smaller variances) than their estimates based on either visual or haptic information alone. Alais and Burr (2004) accounted for the ''ventriloquist effect'' by showing that people combine visual and auditory information in a statistically optimal manner when judging the location of events. Fetsch, DeAngelis, and Angelaki (2010) reviewed evidence that people integrate visual and vestibular heading cues in a manner consistent with optimal integration theory. Wozny, Beierholm, and Shams (2008) showed that people combine visual, auditory, and tactile information optimally when performing a numerosity judgment task.

A second hypothesis studied by cognitive neuroscientists is that training in multisensory environments leads to better learning than training in unisensory environments, and that this advantage holds even when testing is performed in unisensory conditions. Shams and Seitz (2008) argued that unisensory training is less efficient than multisensory training because it is unnatural and thus fails to tap into multisensory learning mechanisms that have evolved to produce optimal behavior in the naturally occurring multisensory environment.

Several investigators have reported experimental results supporting this hypothesis. Seitz, Kim, and Shams (2006) trained subjects to perform a motion detection task based on visual input or based on both visual and auditory input. When tested with visual input alone, subjects trained in the multisensory environment performed significantly better, both on the first day of learning and across all 10 experimental sessions. von Kriegstein and Giraud (2006) trained subjects to recognize voices based on auditory signals or based on auditory signals coupled with visual images of faces. Subjects trained in the multisensory setting performed significantly better even when test trials contained only auditory information. Lehmann and Murray (2005) asked subjects to indicate when an image appeared that had previously been presented during the experiment. Even though sound was never presented when an image recurred, images that previously appeared with congruent sounds (e.g., an image of a bell and the sound ''dong'') were recognized better than those that had previously been presented either without sound or with incongruent sounds (e.g., an image of a bell and the sound ''meow'').

A third hypothesis recently proposed by cognitive neuroscientists is that our neural representations of environmental events are often modality invariant, meaning they are the same (or at least similar) regardless of the sensory modalities through which we perceive those events. From a computational perspective, modality invariance is a desirable property. It will be easier to recognize, reason, and learn about an event if the event has the same representation regardless of the sensory modality through which it is perceived.

Evidence consistent with this hypothesis has been obtained by several researchers. Konkle, Wang, Hayward, and Moore (2009) found that motion aftereffects transfer between vision and touch, suggesting the existence of shared representations of motion. Amedi, Malach, Hendler, Peled, and Zohary (2001) showed that a neural region known as the lateral occipital complex shows similar patterns of activation regardless of whether an object is seen or touched. Quiroga, Kraskov, Koch, and Fried (2009) found single neurons in human brains that respond selectively to the same individual regardless of whether an observer sees a photograph of the individual, reads the name of the individual, or hears the individual's voice.

Lastly, researchers have hypothesized that representations based on different modalities are associated with each other. Consider an observer that sees, but does not hear, an environmental event. A visual representation of that event will be active in the observer's brain, and this representation will often predict or activate an auditory representation of the event even though the event is not heard. From a computational viewpoint, predictions of one modality's sensory representations based on another modality's representations might be useful top-down information leading to faster or more accurate processing in the first modality. These predictions may also help the observer efficiently allocate attention in this modality.

Data consistent with this hypothesis have appeared in the literature. Calvert et al. (1997) reported that viewing facial movements associated with speech (lipreading) leads to activation of auditory cortex in the absence of auditory speech sounds (see also Pekkola et al., 2005; Tanabe, Honda, & Sadato, 2005). Zhou and Fuster (2000) found sustained activity in the primary somatosensory cortex during a delay period after presentation of a visual stimulus previously associated with a tactile stimulus. Sathian, Zangaladze, Hoffman, and Grafton (1997) found that the primary visual cortex (area V1) is active when observers perform a tactile discrimination task involving oriented gratings. Zangaladze, Epstein, Grafton, and Sathian (1999) argued that V1 is crucial for tactile discrimination because disruption of V1 activation using transcranial magnetic stimulation impairs performance on this tactile task.

In summary, this section has reviewed four hypotheses[1] about multisensory perception that will be addressed by the computer simulations reported below:

- Multisensory representations are essential because sensory integration provides significant statistical advantages.
- Training in multisensory environments leads to better learning than training in unisensory environments, even when testing is performed in unisensory conditions.
- Neural representations of events are often modality invariant, meaning they are the same (or at least similar) regardless of the sensory modalities through which the events are perceived.
- Representations based on different modalities are associated with each other.

3. Preliminary remarks regarding the proposed model

To date, there has been relatively little research on how people acquire multisensory representations. An important hypothesis is that multisensory representations, especially representations of complex perceptual events, develop slowly. Unisensory representations develop first, and multisensory representations are acquired later based on statistical correlations among the unisensory representations (Alais, Newell, & Mamassian, 2010). At least in part, this hypothesis is motivated by neuroscientific evidence obtained in physiological investigations of the superior colliculus (SC), a region found in mammalian brains in which sensory integration has often been studied. For example, Wallace and Stein (2001) found that some SC neurons in newborn monkeys respond to signals from multiple sensory modalities, but that these neurons do not synthesize and represent complex multisensory events until later in life. These authors wrote, ''These data, coupled with those from cat, suggest that the capacity to synthesize multisensory information does not simply appear in SC neurons at a prescribed maturational stage but rather develops only after substantial experience with cross-modal cues.''

To this main hypothesis, we add the conjecture that the acquisition of multisensory representations is often accomplished in an unsupervised or task-independent manner. This conjecture is consistent with (but not proved by) recent observations by Lacey, Hall, and Sathian (2010), who found that the performances of subjects performing a shape discrimination task were impaired by changes to objects' task-irrelevant surface properties such as surface texture. The authors concluded that our multisensory representations integrate shape and modality-independent surface properties. If so, this result implies that task-irrelevant features, such as surface texture, are represented in multisensory object representations, thereby suggesting that these representations are acquired in an unsupervised or task-independent manner.

At an intuitive level, the idea that unisensory representations develop first and multisensory representations develop later (in an unsupervised manner) based on statistical correlations among the unisensory features is reasonable and appealing. Nonetheless, its lack of detail makes it difficult to rigorously understand, evaluate, and extend. What predictions, if any, does this hypothesis make about multisensory perception? Does it predict that sensory integration provides significant functional advantages? Does it predict that multisensory training will be better than unisensory training even when testing is conducted in unisensory conditions? Does it predict the existence of modality-invariant multisensory representations? Does it predict the existence of associations among unisensory representations?

Guided by the main hypothesis and conjecture, we propose a model, referred to as the multisensory perception model, of the acquisition of multisensory representations. This model represents a novel approach to modeling the acquisition of multisensory representations, complementary to previous models that explored how acquisition might be algorithmically or neurally implemented. In the traditions of ''ideal observer analysis'' (Barlow, 1961; Geisler, 2004) or ''rational analysis'' (Anderson, 1990; Chater & Oaksford, 1999; Marr, 1982), we consider the abstract computational problem of learning multisensory representations that explain unisensory features in a generative or statistical manner described

below, and show that several aspects of multisensory perception emerge as part of the optimal solution to this learning problem. Although the artificial intelligence, cognitive science, and computational neuroscience literatures contain several computational models (including Bayesian models) of sensory integration (e.g., Abidi & Gonzalez, 1992; Alvarado, Rowland, Stanford, & Stein, 2008; Anastasio, Patton, & Belkacem-Boussaid, 2000; Ernst & Banks, 2002; Hershey & Movellan, 1999; Landy, Maloney, Johnston, & Young, 1995; Massaro, Cohen, Campbell, & Rodriguez, 2001; Pouget, Deneve, & Duhamel, 2002), our model is, to our knowledge, the first attempt focusing on the problem of learning complex multisensory representations from a rational perspective (Anderson, 1990; Chater & Oaksford, 1999; Marr, 1982).

The model can be regarded as an implementation of the main hypothesis and conjecture described above for the purpose of addressing open questions about their meanings, implications, and potential extensions. It is a model of the acquisition of multisensory representations that learns multisensory representations by learning about the statistical correlations among unisensory features. Moreover, it learns in an unsupervised manner.

In unsupervised learning, the data provided to a learner are unlabeled. The goal of the learner is to discover patterns and structure within the data set. There is a dichotomy in the cognitive science and machine learning literatures between parametric and nonparametric unsupervised learning methods. A parametric method uses a fixed representation that does not grow structurally as more data are observed. In contrast, nonparametric methods use representations that are allowed to grow structurally. These methods are advantageous when the goal is to impose as few assumptions as possible. For example, Dirichlet process mixture models are nonparametric models used to cluster data items (Ferguson, 1973; Neal, 2000; Rasmussen, 2000). However, unlike their parametric counterparts such as conventional Gaussian mixture models, they do not make assumptions about the number of clusters from which a set of data items are drawn. Similarly, Indian buffet processes are nonparametric models used to learn latent or hidden features underlying a set of observable variables (Griffiths & Ghahramani, 2005, 2006). Unlike their parametric counterparts such as factor analysis models, they do not make assumptions about the number of latent or hidden features best characterizing the observable variables. In this sense, Bayesian nonparametric techniques are said to ''let the data speak for themselves'' (Blei, Griffiths, & Jordan, 2010).

While both parametric and nonparametric methods have important roles to play in the study of human cognition, Bayesian nonparametric methods are appealing for our purposes because they have both the advantages of probabilistic methods, due to their foundations in Bayesian statistics, and the advantages of flexible representations, due to their nonparametric nature. With respect to representational flexibility, we regard the Bayesian nonparametric approach as an important advance over conventional parametric approaches in which a researcher sets, for instance, the number of latent variables by hand, often in an ad hoc or unprincipled manner (Blei et al., 2010).[2] How can a researcher be sure that the number of latent features should, for example, be exactly 10? Should not the number of latent features be determined by the structure of the task or data set? We also regard the Bayesian nonparametric approach as an advance over modeling approaches that define a set of models, each with a different number of latent features, for instance, and perform ''model comparison'' to

select the best model (see Navarro, Griffiths, Steyvers, & Lee, 2006, for additional discussion of the relationships between Bayesian nonparametric approaches and approaches based on model comparison). Typical model comparison techniques are computationally expensive and, thus, only practical for comparing small numbers of models. How should a researcher pick a small number of models to consider? The Bayesian nonparametric approach eliminates (or at least ameliorates) the problems associated with model comparison.

The proposed model is an extension of a Bayesian nonparametric method known as the Indian buffet process (Griffiths & Ghahramani, 2005, 2006). It ''explains'' the representations arising from individual sensory modalities through the use of a set of latent variables representing multisensory information. Importantly, the number of latent variables is not fixed. Instead, this number is treated as a random variable whose probability distribution is estimated based on the unisensory data. Because the size of the latent multisensory representation is estimated from the observed unisensory data, nonparametric statistical methods are required for inference. As noted by Austerweil and Griffiths (2009), Bayesian nonparametric methods are particularly appropriate for the study of perceptual learning. It is known that people do not use a fixed number of perceptual features (Goldstone, 1998). Instead, people create new features, at least in part, by combining or differentiating existing features. Goldstone (1998) referred to these processes as unitization and differentiation, respectively. Hence, any plausible model of perceptual learning must allow the number of perceptual features to adapt.

The Indian buffet process does not make assumptions about the exact number of latent features underlying a finite set of observed objects, although it does make other assumptions about these features (Griffiths & Ghahramani, 2005, 2006). It assumes that the latent features are binary. Thus, an object either does or does not possess a feature. It also assumes that latent features are statistically independent, meaning that knowledge that an object possesses one feature does not provide information about whether it possesses other features. Lastly, it assumes that the latent features are a finite subset of an unbounded or infinite set of features.

4. Multisensory perception model

We describe the proposed model here in the context of a visual-auditory environment, although we note that the model is equally applicable to other sensory modalities and to any number of modalities (it can also be used when multiple cues arise within a single modality, such as visual stereo, visual motion, and visual shading cues). A coarse schematic of the model is illustrated on the left side of Fig. 1. It contains three sets of nodes or variables corresponding to visual features, auditory features, and multisensory features. The visual and auditory features are statistically dependent. However, they are conditionally independent given values for the multisensory features. The values of the visual features are observed when an object is viewed. When an object is not viewed, the visual features are latent, and their distributions can be inferred. Similarly, the values of the auditory features are observed when an object is heard. Otherwise, the auditory features are latent, and their
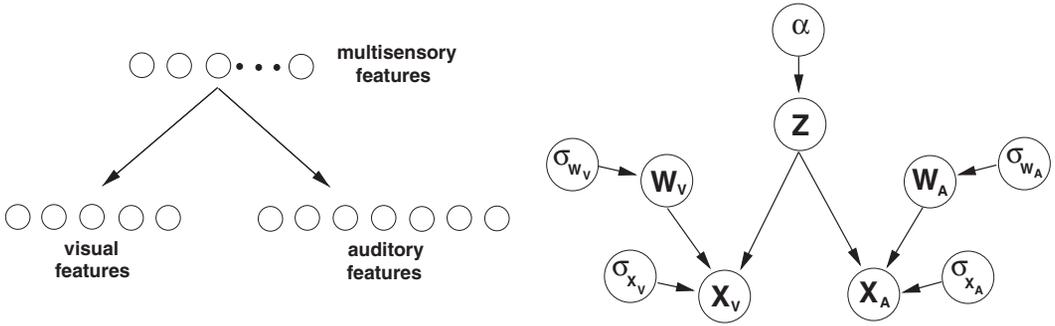
Fig. 1. (Left) A coarse schematic of the multisensory perception model. (Right) A Bayesian network representation of the multisensory perception model.

distributions can be inferred. The multisensory features are always latent variables. Whereas the numbers of visual and auditory features are fixed, the number of multisensory features is not. Consistent with the nonparametric approach, this number is a random variable whose distribution is inferred from the data.

Formally, the model is a straightforward extension of the Indian buffet process (Griffiths & Ghahramani, 2005, 2006). A detailed graphical representation of the model is shown on the right side of Fig. 1. An important goal of the model is to find a set of latent multisensory features, denoted Z, ''explaining'' a set of observed visual and auditory features, denoted X_V and X_A, respectively. Assume that a learner both sees and hears a number of objects. Let Z be a binary multisensory feature ownership matrix, where $Z_{ij} = 1$ indicates that object i possesses multisensory feature j. Let X_V and X_A be real-valued visual and auditory feature matrices, respectively (e.g., $X_{V_{ij}}$ is the value of visual feature j for object i). The problem of inferring Z from X_V and X_A can be solved via Bayes' rule:

$$p(Z|X_V, X_A) = \frac{p(X_V|Z)p(X_A|Z)p(Z)}{\sum_{Z'} p(X_V|Z')p(X_A|Z')p(Z')}, \tag{1}$$

where $p(Z)$ is the prior probability of the multisensory feature ownership matrix, and $p(X_V|Z)$ and $p(X_A|Z)$ are the likelihoods of the observed visual and auditory feature matrices, respectively, given the multisensory features. We now describe the prior and likelihood distributions.

The multisensory feature ownership matrix is assigned a Bayesian nonparametric prior distribution known as the Indian buffet process (Griffiths & Ghahramani, 2005, 2006). It can be interpreted as a probability distribution over feature ownership matrices with an unbounded (infinite) number of features. The distribution is written as:

$$p(Z) = \frac{\alpha^K}{\prod_{h=1}^{2^N-1} K_h!} \exp\{-\alpha H_N\} \prod_{k=1}^{K} \frac{(N - m_k)!(m_k - 1)!}{N!}, \tag{2}$$

where N is the number of objects, K is the number of multisensory features, K_h is the number of features with history h (the history of a feature is the matrix column for that feature interpreted as a binary number), H_N is the Nth harmonic number, m_k is the number of objects

with feature k, and α is a variable influencing the number of features. (We did not set α to a fixed value in our simulations described below. To make the model more flexible, α was assigned a vague prior distribution, and its posterior distribution was inferred.)

The visual and auditory likelihoods are each based on a linear-Gaussian model. Let z_i be the multisensory feature values for object i, and let $x_{i\beta}$ be the feature values for object i where β is set to either V or A depending on whether we are referring to visual or auditory features. Then $x_{i\beta}$ is drawn from a Gaussian distribution whose mean is a linear function of the multisensory features, z_iW_β, and whose covariance matrix equals $\sigma^2_{X_\beta}I$, where W_β is a weight matrix (the weight matrices themselves are drawn from zero-mean Gaussian distributions with covariance $\sigma^2_{W_\beta}I$). Given these assumptions, the likelihood for a feature matrix is:

$$p(X_\beta|Z, W_\beta, \sigma^2_{X_\beta}) = \frac{1}{(2\pi\sigma^2_{X_\beta})^{ND_\beta/2}}\exp\{-\frac{1}{2\sigma^2_{X_\beta}}\mathrm{tr}((X_\beta - ZW_\beta)^T(X_\beta - ZW_\beta))\}, \quad (3)$$

where D_β is the dimensionality of X_β, and $\mathrm{tr}(\cdot)$ denotes the trace operator.

Exact inference in the model is computationally intractable and, thus, approximate inference must be performed using Markov chain Monte Carlo (MCMC) sampling methods (Gelman, Carlin, Stern, & Rubin, 1995; Gilks, Richardson, & Spiegelhalter, 1996). In our simulations, we used the MCMC sampling algorithms of Griffiths and Ghahramani (2005).

5. Synthetic data set

This section applies the multisensory perception model to a synthetic data set. Our goal is to illustrate for the reader that the model performs sensory integration in a sensible manner. Toward this goal, we consider an experiment by de Gelder and Vroomen (2000). This experiment is useful for our purposes because it demonstrates important properties of human sensory integration in a straightforward and clear fashion.

de Gelder and Vroomen (2000) studied how people integrate visual and auditory information when judging the emotional content of facial expressions. They created a continuum of images of a person's face ranging from a happy face at one end to a sad face at the other end. They also created a continuum of voices ranging from a happy voice to a sad voice. On each trial, subjects saw a face and heard a voice, and judged whether the facial expression was happy or sad. The results show that the voice influenced subjects' categorizations. When a happy voice was paired with a facial expression, subjects were more likely to judge the face as happy than when the face was paired with a neutral or sad voice. Similarly, subjects were more likely to judge a face as sad when it was paired with a sad voice than when it was paired with a neutral or happy voice. The authors referred to this result as an ''emotional McGurk effect'' (de Gelder & Vroomen, 2003).

Does the multisensory perception model show an analogous effect? To address this question, we created a synthetic set of stimuli that is meant to be similar in spirit to the stimuli used by de Gelder and Vroomen (2000). We defined a line in a two-dimensional space, and

we defined nine categories along this line. The category prototypes were located at [4,4],[3,3],…,[−4,−4]. A normal distribution was centered at each prototype, where a distribution's covariance matrix was $\sigma^2 I$ ($\sigma^2 = 0.3$). Thirty samples were collected from each distribution. Twenty were used during training, and 10 were used during testing.

For ease of explanation, we take some terminological liberties. First, we assign emotions to categories. Categories whose associated distributions are centered at positive coordinates are referred to as ''happy,'' the category whose associated distribution is centered at [0, 0] is referred to as ''neutral,'' and categories whose associated distributions are centered at negative coordinates are referred to as ''sad.'' In addition, the multisensory perception model had two sets of input variables, which we will refer to as ''visual'' and ''auditory'' features.

During training, input variables were consistent, meaning that visual and auditory features were sampled from the same distribution or category. A single MCMC chain was simulated. The chain was run for 500 iterations. The first 400 iterations were discarded as burn-in. In simulations of additional chains, we found that the results reported here are typical.

Our simulations included three test conditions. In these conditions, the model was exposed to data items potentially containing sensory conflicts. In all conditions, visual features could come from any of the nine categories. In the first test condition, auditory features always came from a happy category (the category centered at [1, 1]). In the second test condition, auditory features always came from the neutral category. Auditory features always came from the sad category (the category centered at [−1,−1]) in the third test condition.

Evaluating the model's performance on test items presents unique challenges. Although it is reasonable to sample variables' values, and thus estimate variables' distributions, on the basis of training items, models are not meant to learn from test items. Consequently, we could not run the MCMC sampler on the model using the test items to evaluate the model's categorization performance. Doing so would erase the distinction between training and test data items.

Instead, we proceeded as follows. Consider the latent feature representations obtained on iteration i of the MCMC sampler when the model was trained on the training data. There is one such representation for each training item. These are the latent representations with non-zero probability based solely on iteration i. Let $\mathcal{L}_i$ denote this set of representations. For each test item in each test condition, we searched $\mathcal{L}_i$ to find the latent representation that was most probable given the item. The item was then categorized on the basis of this latent representation. Specifically, the item was assigned to category j if its most probable latent representation was associated with category j on training iteration i. Categorization performances were averaged across iterations.

The results are shown in Fig. 2. The horizontal axis plots the category from which a test item's visual features were sampled, and the vertical axis plots the probability that the model classified the test item as sad. Data points indicated by diamonds correspond to the first test condition (auditory features were sampled from the happy category centered at [1, 1]), points indicated by circles correspond to the second test condition (auditory features were sampled from the neutral category centered at [0, 0]), and points indicated by squares

correspond to the third condition (auditory features were sampled from the sad category centered at $[-1,-1]$).

These results demonstrate that the multisensory perception model integrates information from two sources in a sensible manner. The model was more likely to judge a test item as happy when visual features were paired with happy auditory features than when paired with neutral or sad auditory features. Similarly, the model was more likely to judge a test item as sad when visual features were paired with sad auditory features than when paired with neutral or happy auditory features. In this sense, the model replicates the ''emotional McGurk effect.''

6. Visual-auditory data set

We applied the multisensory perception model to a real-world visual-auditory data set known as Tulips1 (Movellan, 1995). Twelve people (nine adult males, three adult females) were videotaped while uttering the first four digits of English twice.

In each video frame, the image of a speaker's mouth was processed to extract six visual features: the width and height of the outer corners of the mouth, the width and height of the
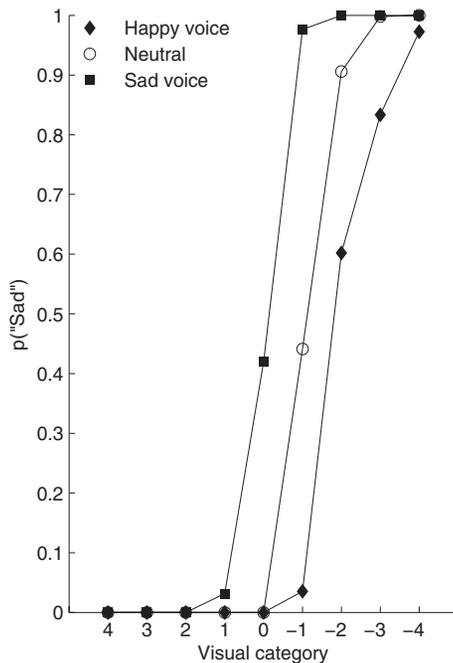

Fig. 2. The performance of the multisensory perception model on the first synthetic data set. The horizontal axis plots the category from which a test item's ''visual'' features were sampled, and the vertical axis plots the probability that the model classified the test item as ''sad.'' Data points indicated by diamonds correspond to the first test condition (auditory features were sampled from a ''happy'' category), points indicated by circles correspond to the second test condition (auditory features were sampled from the neutral category), and points indicated by squares correspond to the third condition (auditory features were sampled from a ''sad'' category).

inner corners of the mouth, and the heights of the upper and lower lips. The auditory signal corresponding to a frame was processed to extract 26 features: 12 cepstral coefficients (these are the coefficients of the Fourier transform representation of the log magnitude spectrum), 1 log-power, 12 cepstral coefficient derivatives, and 1 log-power derivative. Because speech utterances had different durations, we sampled six frames for each utterance spanning the entire duration of the utterance in a uniform manner. In summary, each data item contained values for 36 visual features (6 frames × 6 visual features per frame) and 156 auditory features (6 frames × 26 auditory features per frame).

Training and test sets were created as follows. For the first eight speakers, one utterance of each digit was used for training and the other utterance was used for testing. For the remaining speakers, both utterances were used for training. Thus, the training set contained 16 data items for each digit, and the test set contained 8 data items for each digit.

To understand better the performance of the multisensory perception model, we also consider the performances of two other models. The vision-only model is identical to the multisensory model except that it contains only two sets of variables corresponding to visual and latent features. When applied to the Tulips1 data set, it received only the visual features. Similarly, the audition-only model contains only two sets of variables corresponding to auditory and latent features. It received only the auditory features from the data set.

A single MCMC chain of each model was simulated. The chain was run for 5,000 iterations, where the first 3,000 iterations were discarded as burn-in. To reduce correlations among variables at nearby iterations, the remaining iterations were thinned to every 10th iteration (i.e., only variable values at every 10th iteration were retained). Thus, the results below are based on 200 iterations. Simulations of additional chains produced nearly identical results as those reported here.

6.1. Posterior distributions over latent features

Recall that the number of latent features in each model is not fixed a priori. Instead, it is a random variable whose distribution is inferred from the training data. The three graphs in Fig. 3 show the distributions of the numbers of latent features in the vision-only, audition-only, and multisensory models. The vision-only model used relatively few latent features, the audition-only model used more latent features, and the multisensory model used the most latent features. This result confirms that the models are highly flexible. Their nonparametric nature allows them to adapt their representational capacities based on the complexities of their data sets. Interestingly, the multisensory model learned a compact set of latent features: The number of features it acquired was always less than the sum of the number of features acquired by the vision-only and audition-only models.

6.2. Categorization performances

We evaluated each model's ability to categorize the speech utterances as instances of one of the first four digits in English based upon its latent feature representations. At each iteration of an MCMC chain, a model sampled a latent feature representation for each data item
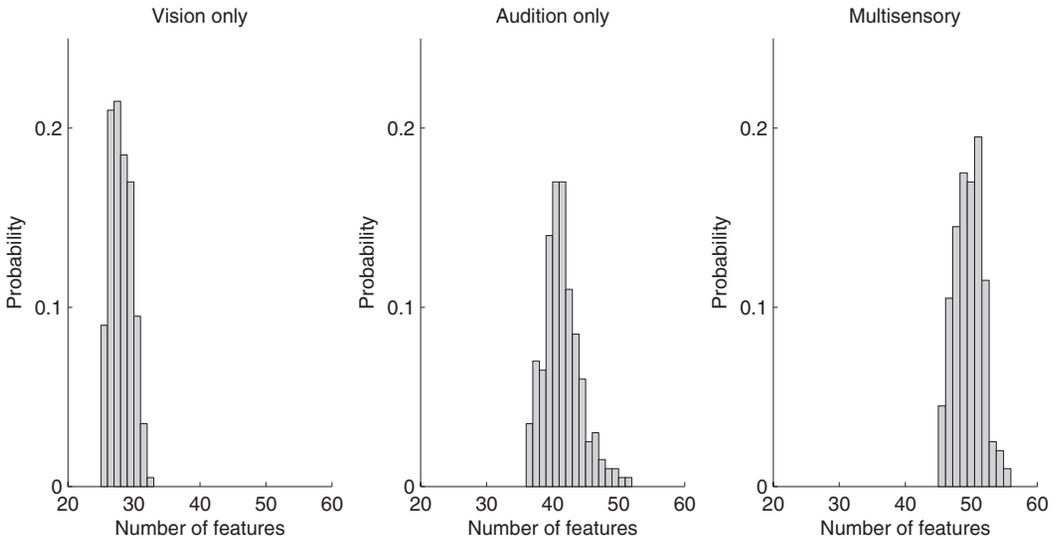
Fig. 3. The distributions of the numbers of latent features in the vision-only (left), audition-only (middle), and multisensory (right) perception models.

in the training set. Using these representations, we performed *k*-means clustering with four cluster centers. We then performed an exhaustive search of assignments of clusters to English digits (e.g., cluster A → digit 3, cluster B → digit 1, etc.) to find the assignment producing the best categorization performance. Performances were averaged across iterations of a chain.

The results are shown in the leftmost graph of Fig. 4. The horizontal axis gives the model, and the vertical axis plots the percent of data items in the training set that were correctly classified (error bars indicate the standard deviations of these percents across iterations of an MCMC chain). As expected, the vision-only model showed the worst performance, the audition-only model showed better performance, and the multisensory model showed the best performance (based on two-tailed *t* tests, the differences between performances of the multisensory model and each of the other models are statistically significant at the $p < .05$ level).

It is possible that the multisensory model showed the best performance solely due to the fact that it received both visual and auditory features and, thus, received a richer set of inputs than either the vision-only or audition-only models. To evaluate this possibility, we simulated a model, referred to as a ''mixed'' model, that resembled the multisensory model in the sense that it received both visual and auditory features. However, for the mixed model, these features were not segregated into separate input streams. Instead, the mixed model contained a set of latent features that received inputs from a set of undifferentiated perceptual features, namely a concatenation of the visual and auditory features. The results for the mixed model on the training set are also shown in the leftmost graph of Fig. 4. The mixed model showed significantly poorer performance than the multisensory model, thus
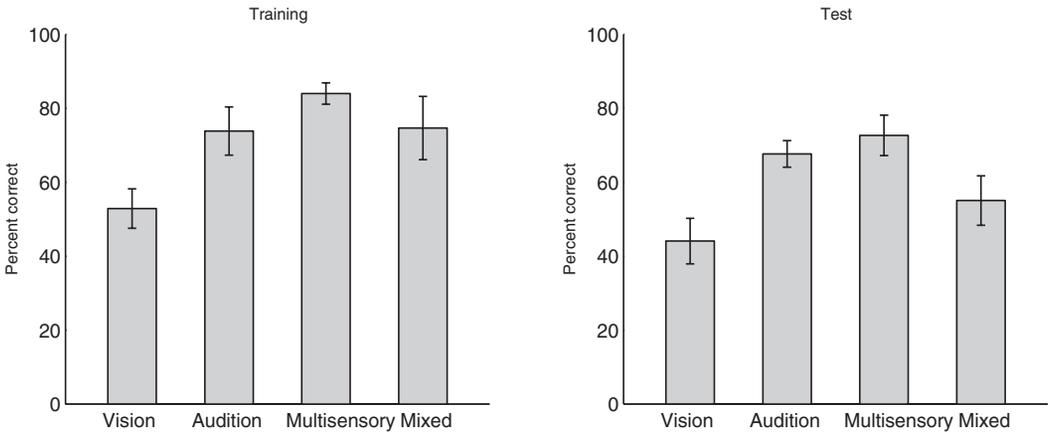
Fig. 4. Categorization performances of the vision-only, audition-only, multisensory, and mixed models on the training set (left), and on the test set (right). The horizontal axis of each graph gives the model, and the vertical axis plots the percent of data items correctly classified (error bars indicate the standard deviations of these percents across iterations of an MCMC chain).

suggesting that the multisensory model benefited from independently accounting for visual and auditory features (albeit at the expense of additional variables).

This analysis was repeated using the data items in the test set. For a given model, let $\mathcal{L}_i$ denote the latent feature representations obtained on iteration i of the MCMC sampler when the model was trained on the training data. Recall that there is one such representation for each training item, and these are the latent representations with nonzero probability based solely on iteration i. For each data item in the test set, we searched $\mathcal{L}_i$ to find a latent representation that was most probable given the item. This was repeated for every item in the test set. Using these representations, the analysis of the test set is identical to the analysis of the training set described above: Latent representations were clustered using k-means clustering, and a search of assignments of clusters to digits was performed to find the assignment producing the best categorization performance. Performances were averaged across iterations.

The results are shown in the rightmost graph of Fig. 4. Again, the multisensory model showed the best performance.

In summary, the multisensory perception model showed the best categorization performance on both training and test data sets. We conclude that its superior performance is due to both its rich set of inputs (it receives both visual and auditory features) and due to its internal structure (visual and auditory features are segregated perceptual streams). Clearly, this model received the statistical benefits of sensory integration.

6.3. Multisensory versus unisensory training

Above, we reviewed experimental evidence that training in multisensory environments leads to better learning than training in unisensory environments, and that this advantage

occurs even when testing is conducted under unisensory conditions. Does the multisensory perception model predict the superiority of multisensory training?

To study this question, we compared the categorization performances of the multisensory and vision-only models when visual features were the only inputs to these models, and compared the performances of the multisensory and audition-only models when auditory features were the only inputs. For ease of explanation, we first consider the multisensory and vision-only models when visual features were the only inputs.

As above, let $\mathcal{L}_i$ denote the set of multisensory feature representations obtained on iteration i of the MCMC sampler when the model was trained on the training data. Again, these are the latent or multisensory representations with nonzero probability based solely on iteration i. For each data item in the training set, we calculated the most probable multisensory representation based solely on an item's visual features. A set of most probable representations is formed when all training items are taken into account. Using this set, we categorized the data items in the same manner as above: Multisensory representations were clustered using k-means clustering, and a search of assignments of clusters to digits was performed to find the assignment producing the best categorization performance.

For the vision-only model, categorization of training items was performed in a similar manner. At each iteration of an MCMC chain, the vision-only model sampled a latent feature representation for each data item in the training set. Categorization via k-means clustering was performed on the basis of these representations.

The results are shown on the left side of the left graph in Fig. 5. The horizontal axis labels the model, and the vertical axis plots the proportion of data items a model categorized correctly (error bars plot the standard deviations of these proportions across iterations of an MCMC chain). Clearly, the multisensory perception model outperformed the vision-only model, even though both models were evaluated solely on the basis of training items' visual features.

For data items in the test set, the categorization performances of the multisensory and vision-only models were calculated in analogous ways. For each model, the most probable latent representations were found at each iteration based on the items' visual features. Categorization was performed on the basis of these representations. The results are shown on the right side of the left graph in Fig. 5. Again, the multisensory perception model performed better than the vision-only model.

Analogous computations were carried out to compute the categorization performances of the multisensory and audition-only models when auditory features were the only inputs to these models. The results are shown in the right graph of Fig. 5. For training items, the multisensory model outperformed the audition-only model. For test items, however, the two models showed similar categorization performances.

In summary, the multisensory perception model performed better than the unisensory models. This result was obtained even though the multisensory model was trained in multisensory conditions but evaluated in unisensory conditions. The result was strongest when models were evaluated on the basis of visual features only. This is expected because auditory features contain more information about the identity of speech utterances than visual
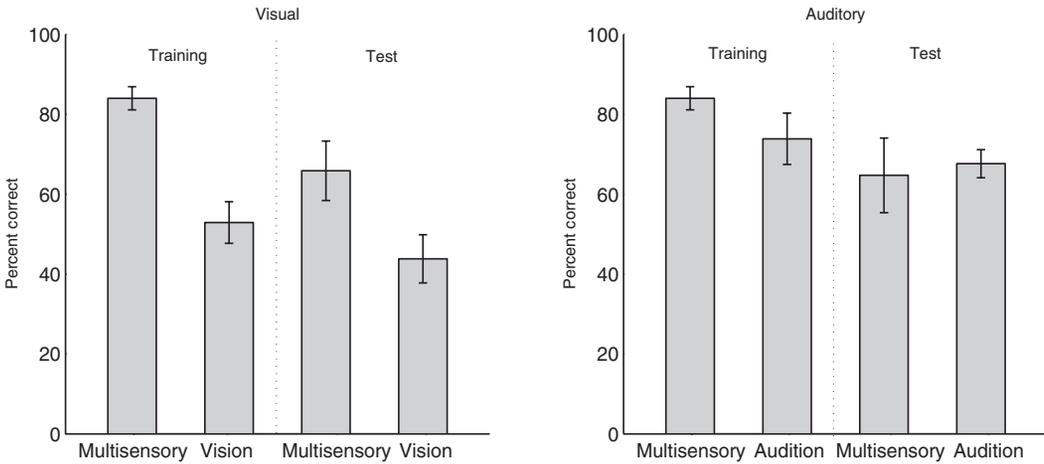
Fig. 5. (Left) Categorization performances of the multisensory and vision-only models on the training (left side) and test (right side) data items when visual features were the only inputs to these models. The horizontal axis labels the model, and the vertical axis plots the proportion of data items a model categorized correctly (error bars plot the standard deviations of these proportions across iterations of an MCMC chain). (Right) Categorization performances of the multisensory and audition-only models on the training and test data items when auditory features were the only inputs to these models.

features, and thus visual-auditory training should have its largest impact when evaluation is conducted solely with visual features. Clearly, the multisensory perception model predicts that training in multisensory environments should lead to better learning than training in unisensory environments, and that this advantage should occur even when evaluation is conducted under unisenseory conditions.

6.4. Modality invariance

As discussed above, neural representations of objects are often modality invariant (Amedi et al., 2001; Konkle et al., 2009; Quiroga et al., 2009). That is, the same (or at least similar) neural representations arise regardless of the modality through which an object is sensed. Does the multisensory perception model show this same property?

We investigated this question as follows. Once again, let $\mathscr{L}_i$ denote the set of multisensory feature representations obtained on iteration i of the MCMC sampler when the model was trained on the training data. These are the multisensory representations with nonzero probability based solely on iteration i. For each data item in the training set, we calculated the probability distribution of the multisensory representation given an item's visual features, and the distribution of the multisensory representation given an item's auditory features where $\mathscr{L}_i$ was the set of possible multisensory representations. When all training items are taken into account, these distributions are denoted $p(Z|X_V)$ and $p(Z|X_A)$, respectively. We then calculated the Battacharyya distance[3] between $p(Z|X_V)$ and $p(Z|X_A)$. On every iteration, this distance was zero.

We repeated this analysis using the data items in the test set. Again, we computed $p(Z|X_V)$ and $p(Z|X_A)$ where X_V and X_A refer to the visual and auditory features of test items, and where $\mathscr{L}_i$ is the set of possible multisensory representations. The Battacharyya distances between $p(Z|X_V)$ and $p(Z|X_A)$ are always small values—the distribution of these distances has values of 1.51, 1.55, and 1.68 as its 25th, 50th, and 75th percentiles, respectively. By way of comparison, we also computed the distance between $p(Z|X_A)$ and a uniform distribution over multisensory representations. The distribution of these distances has values of 3.49, 7.83, and 19.04 as its 25th, 50th, and 75th percentiles.

In summary, both training and test sets suggest that the multisensory perception model did indeed acquire modality-invariant representations. Its latent multisensory features had the same or similar distributions regardless of whether a speech utterance was seen or heard.

6.5. Predicting sensory representations in missing modalities

Above, we reviewed evidence of activity in people's auditory cortices when they viewed speech utterances but did not hear those utterances (Calvert et al., 1997; Pekkola et al., 2005). This result is consistent with the hypothesis that sensory representations in one modality can predict or activate representations in other modalities. Does the multisensory perception model support the ability to use sensory representations in one modality to predict representations in other modalities?

This question was studied using the data items in the test set. Let $\mathscr{V}$ and $\mathscr{A}$ denote the sets of visual and auditory feature representations for the data items in the training set. Once again, let $\mathscr{L}_i$ denote the set of multisensory representations obtained on iteration i of the MCMC sampler when the model was trained on the training data. For each test item, we computed the probability distribution of an auditory representation given a test item's visual features. This was accomplished by first calculating a conditional joint distribution over both multisensory and auditory representations, and then by marginalizing over the multisensory representations where the set of possible auditory and multisensory representations were given by $\mathscr{A}$ and $\mathscr{L}_i$. Analogous computations were carried out to compute the distribution of a visual representation given an item's auditory features.

Representative results are shown in Fig. 6. Four test items (items 1, 12, 24, and 28) were selected at random with the constraint that one item corresponded to each spoken digit. The four graphs in the top row of the figure show the distributions of the visual representations given the auditory features of the test items. More precisely, the graphs show that when presented with the auditory features corresponding to one of the digits, the model's distribution of visual representations was tightly peaked at a representation corresponding to the same digit. The four graphs in the bottom row show analogous results for distributions of auditory representations given test items' visual features.

In summary, the multisensory perception model learns to associate unisensory representations from different modalities. It successfully predicts representations from missing modalities based on features from observed modalities.
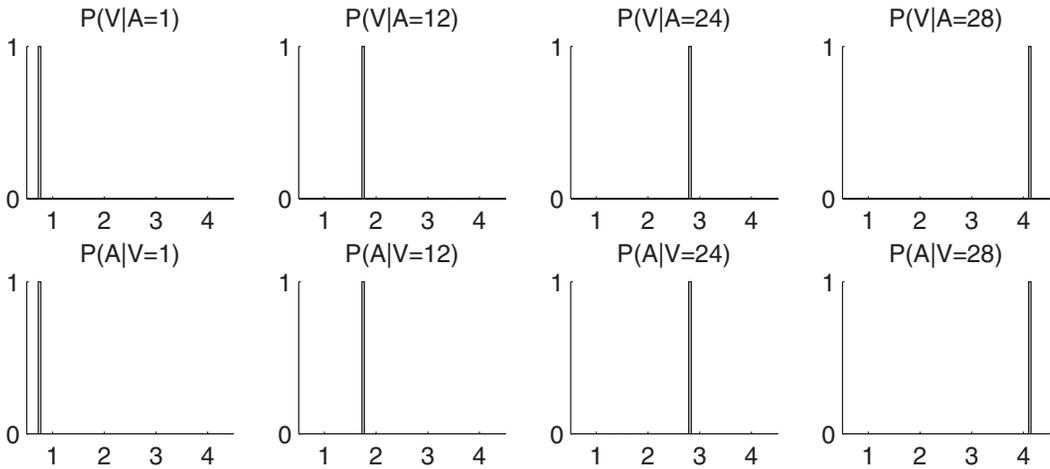
Fig. 6. Graphs in the top row demonstrate that when presented with auditory features of a test item corresponding to one of the digits, the multisensory perception model's distribution of visual representations was tightly peaked at a representation corresponding to the same digit. Graphs in the bottom row show analogous results for distributions of auditory representations given test items' visual features.

7. Visual-haptic data set

This section reports the results of applying the multisensory perception model to an experimental data set collected by Cooke, Jäkel, Wallraven, and Bülthoff (2007). As discussed below, a motivation for this application is that it permits us to interpret the multisensory features acquired by the model.

Cooke et al. (2007) created novel, three-dimensional objects that varied parametrically in shape (macrogeometry) and texture (microgeometry). Subjects performed a similarity rating task in which they repeatedly rated the similarity of pairs of objects. One group of subjects based their similarity judgments on their visual percepts, another group based their judgments on their haptic percepts, and a third group based their judgments on both visual and haptic percepts. Similarity ratings were analyzed using multidimensional scaling. This analysis revealed that people's similarity judgments were strongly dominated by shape when objects were seen. However, shape and texture were roughly equally important in determining similarity when objects were touched or when they were both seen and touched. The authors concluded that the perceptual modality used to interact with objects affects the mental representations used for similarity judgments and categorization.

Here, we apply the multisensory perception model to the experimental data collected by Cooke et al. (2007). Doing so, however, requires a small modification of the model. In the simulations reported above, the observable data used to train a model consisted of unisensory perceptual features. The data collected by Cooke et al. do not consist of perceptual features. Rather they consists of pairwise similarity ratings. The changes needed to modify the multisensory perception model so that it could be trained with similarity data were described by Navarro and Griffiths (2008) in the context of the Indian buffet process. In brief, Navarro

and Griffiths adapted the Indian buffet process so that it can be interpreted as a Bayesian formulation of additive clustering, a technique for inferring the features of a set of stimuli from their similarities (Shepard & Arabie, 1979). Importantly for our purposes, the formulation uses similarity ratings to infer one or more latent or hidden features that ''explain'' these ratings. Associated with each feature is a non-negative weight quantifying the saliency of that feature. [4]

We trained four models using the experimental data of Cooke et al. (2007), a vision-only model, a haptic-only model, and two multisensory models. The vision-only model had two sets of variables. These corresponded to latent features and subjects' average similarity ratings when objects were seen. The haptic-only model also had two sets of variables. These corresponded to latent features and subjects' average ratings when objects were touched. The first multisensory model (Multisensory-1) had two sets of variables corresponding to latent features and subjects' average ratings when objects were both seen and touched. In contrast, the second multisensory model (Multisensory-2) did not receive the data from the multisensory experimental condition. Instead, it received the data from the two unisensory conditions. This model had three sets of variables corresponding to latent features, subjects' ratings when objects were seen, and subjects' ratings when objects were touched. Although the model received only the data from unisensory conditions, we predicted that it would infer latent features resembling those of the model receiving data from the multisensory experimental condition.

A single MCMC chain of each model was simulated. The chain was run for 21,000 iterations. The first 1,000 iterations were discarded as burn-in. To reduce correlations among variables at nearby iterations, the remaining iterations were thinned to every 10th iteration (i.e., only variable values at every 10th iteration were retained). Thus, the results below are based on 2,000 iterations.

The results are shown in Fig. 7. Each panel illustrates the 25 objects used in the experiment of Cooke et al. (2007). Objects vary systematically in texture properties along the horizontal axis, and they vary in shape properties along the vertical axis. The four panels correspond to the four models that we implemented. Each panel illustrates properties of the latent features acquired by its corresponding model. Let $\mathscr{L}_i$ denote the set of latent features acquired by a model on iteration i of the MCMC sampler when the model was trained on the training data. In Fig. 7, we only consider ''stable'' features, meaning those that existed in $\mathscr{L}_i$ for at least 25% of the iterations.

The upper-left panel plots the results for the vision-only model. This model acquired two stable latent features. One feature was possessed by objects 1–15. These objects occupy the bottom three rows of the panel (region enclosed by the solid line). The other feature was possessed by objects 12–25, occupying nearly all of the top three rows (region enclosed by the dotted line). In other words, these features were ''on'' for different values of the shape parameter (vertical axis). The first feature was on when the shape parameter had a small value, and the second feature was on when the parameter had a large value. Clearly, subjects' similarity ratings when objects were seen were based on shape properties. The weight values associated with each of the features are given at the bottom of the panel.
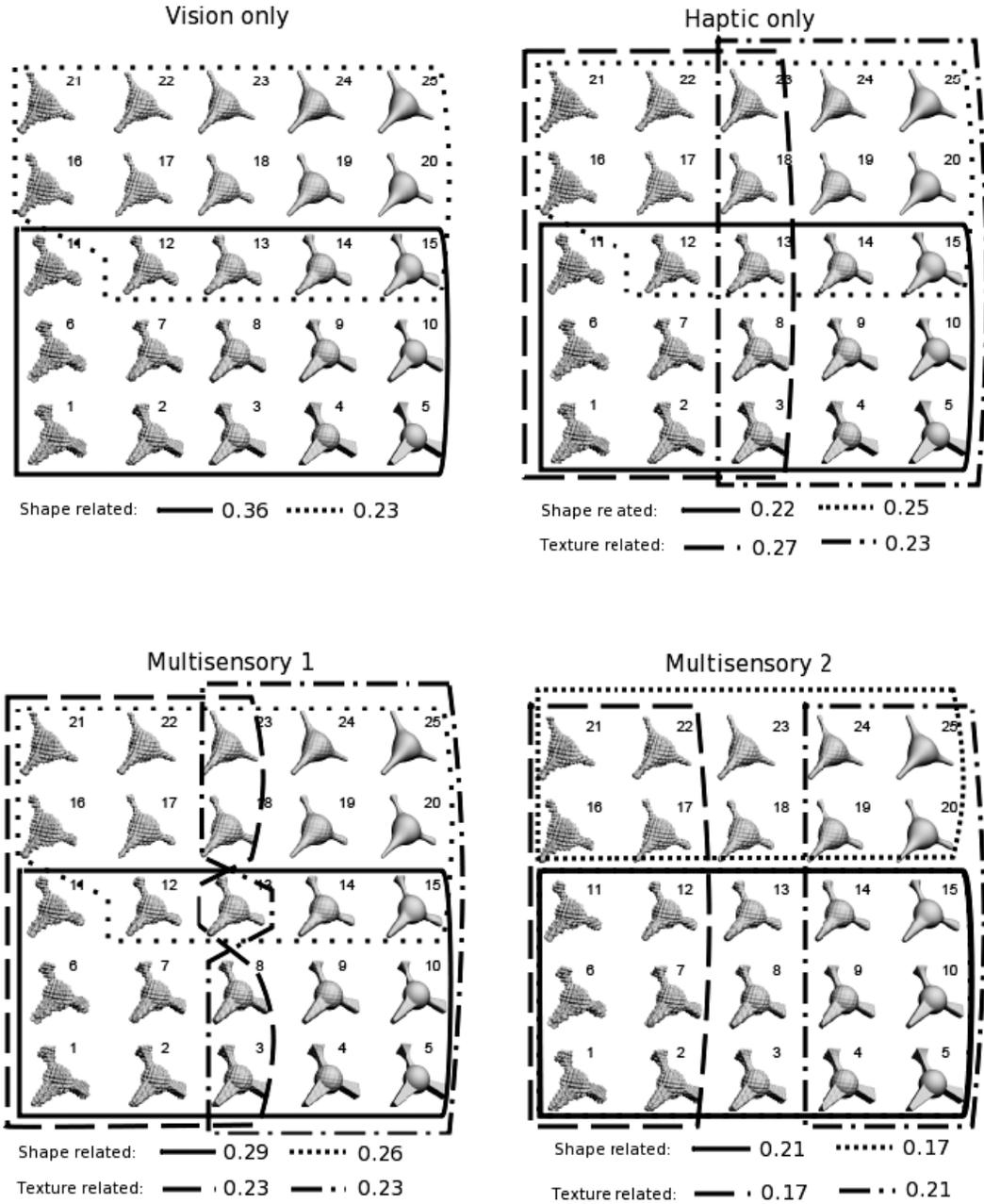
Fig. 7. Each panel illustrates the 25 objects used in the experiment of Cooke et al. (2007). Objects vary systematically in texture properties along the horizontal axis, and they vary in shape properties along the vertical axis. The four panels correspond to the four models that were implemented. Each panel illustrates properties of the latent features acquired by its corresponding model. See text for further explanation (figure of Cooke et al., 2007, adapted with permission from Elsevier).

The upper-right panel plots the results for the haptic-only model. In this case, four latent features were acquired. Two of the features can be characterized as ''shape'' features. They are identical to the features acquired by the vision-only model. One feature was on when the shape parameter had a small value, and the other feature was on when the parameter had a large value. The remaining two features can be characterized as ''texture'' features. One feature was on when the texture parameter (horizontal axis) had a small value, and the other feature was on when the parameter had a large value. In contrast to the vision-only case, subjects' similarity ratings when objects were touched seem to have been based on both shape and texture properties.

As illustrated in the bottom-left panel, the results for the Multisensory-1 model are nearly identical to those of the haptic-only model. Four latent features were acquired. Two of the features can be characterized as shape features, and the remaining two can be characterized as texture features. When objects were both seen and touched, subjects' similarity ratings were based on both shape and texture properties.

The bottom-right panel shows the results for the Multisensory-2 model. Recall that the Multisensory-1 model received one set of multisensory similarity ratings as input, whereas the Multisensory-2 model received two sets of unisensory similarity ratings. Despite this difference, the two models acquired nearly identical latent features. The Multisensory-2 model acquired four latent features, two shape features, and two texture features. (Because the Multisensory-2 model received two sets of similarity ratings, it had two weights associated with each latent feature. The values of these weights were averaged, and these average values are shown at the bottom of the panel.)

In summary, the features acquired by the models studied here are in agreement with the analyses of Cooke et al. (2007). Subjects' similarity ratings seem to be based on shape properties when objects are viewed, and they seem to be based on both shape and texture properties when objects are touched or seen and touched. Moreover, the multisensory perception model (Multisensory-2) acquired multisensory latent features from two unisensory data sets that were nearly identical to the features acquired directly from a single multisensory data set (Multisensory-1). This result suggests that the model acquired multisensory representations that closely resemble those used by human subjects when judging the similarity of objects in a multisensory environment.

8. Discussion

How do people learn multisensory representations, and what consequences do these representations have for perceptual performance? We addressed this question by performing a rational analysis of the problem of learning multisensory representations that optimally explain unisensory features detected by individual modalities (Anderson, 1990; Marr, 1982). This analysis indicates that several aspects of multisensory perception emerge as part of the optimal solution to the problem of learning to represent complex multisensory environments. On the basis of our results, we argue that multisensory representations should (a) be adaptable and compact, (b) integrate information from multiple modalities in such a

way that leads to superior performances in, for example, categorization tasks, (c) lead to better learning than representations acquired in unisensory environments, even when testing is conducted in unisensory conditions, (d) be modality invariant, and (e) support the prediction of ''missing'' sensory representations in modalities when the input to those modalities is absent.

An assumption of the research reported here is that multisensory representations do indeed exist in people's minds and brains. The question of whether behavioral phenomena are best accounted for by postulating the existence of a multisensory representation versus the existence of multiple modality-specific representations has been actively studied. With regard to interactions between visually and haptically derived representations of objects, Lacey, Campbell, and Sathian (2007) concluded that the weight of evidence suggests the existence of a multisensory representation. We concur with their conclusion, and we speculate that multisensory representations are commonplace in biological organisms.

A second assumption of our work is that learning involves modifications to multisensory representations. There are alternatives to this possibility, namely that learning involves modifications to unisensory representations or to connections among unisensory representations. As pointed out by Shams and Seitz (2008), these possibilities are not mutually exclusive. It may be that learning involves all three types of changes. If so, then the work presented here is incomplete due to its exclusive focus on the acquisition of multisensory representations.

We regard the rational analysis performed here as a useful early step toward understanding the acquisition of multisensory representations. Future work will need to address a number of open issues. A prediction of our work is that there is a connection between data complexity and memory code length, and that Bayesian nonparametric models provide a novel method for accounting for this connection. An unusual property of the proposed model is that it adapts the size of its multisensory representation based on the complexity of the unisensory data it receives. This can be understood using intuitions from information theory (Cover & Thomas, 1991). We expect that unisensory data that are highly regular (e.g., highly predictable or low entropy) are mentally represented or ''explained'' with a compact code, such as a small number of multisensory features. In contrast, unisensory data with fewer statistical regularities should be represented by longer codes, such as a large number of multisensory features. Future work will need to test these predictions.

Future work should also include computational extensions to the proposed model. The model currently learns in an unsupervised manner, but it could be extended to a supervised setting. If, in addition to explaining a data item's unisensory features, the multisensory perception model was required to also explain an item's target response, then the model would be applicable to supervised situations. For example, a model could receive as input each item's visual features, auditory features, and category label. The model would learn a multisensory representation that explains all three inputs. Such a model would possess several interesting properties because the multisensory representations acquired by the model would be shaped by the target responses. If the target responses were category labels, then the model would be constrained to acquire similar multisensory representations for data items belonging to the same category. That is, the representations would be task specific,

not task independent. Furthermore, when the target responses are not observed, the model could compute a probability distribution over possible target responses given a test item's unisensory features. Thus, the model could act as a classifier, regressor, or associative memory.

As a second possible extension, the multisensory perception model learns multisensory features at a single level of abstraction, but this representation could be extended to a hierarchy. Courville, Eck, and Bengio (2009) showed how an Indian buffet process can be extended to two or more layers of latent variables. Each layer defines a distribution over the latent variables of the layer below via a noisy-or mechanism (Neapolitan, 2004). They showed that this hierarchical model often outperforms a single-level Indian buffet process in the sense that it achieves higher likelihoods on test data items and more compact latent representations. Similarly, Adams, Wallach, and Ghahramani (2010) developed a hierarchical model, referred to as the Cascading Indian Buffet Process, that learns a layered, directed belief network of latent variables that explain the observable variables. An interesting feature of this model is that it provides a prior distribution on the structure of belief networks that is unbounded in both depth and width, yet allows tractable inference. Applying the hierarchical extensions of either Courville et al. (2009) or Adams et al. (2010) to the multisensory perception model should be straightforward, and it may provide insights into multisensory features at multiple levels of abstraction.

Finally, the model currently does not include a notion of time. Future work might consider a model version that is sensitive to temporal dependencies in sensory data (Williamson, Orbanz, & Ghahramani, 2010) or a version whose inferential processes are time dependent (as in sequential Monte Carlo methods). These modifications would allow the model to potentially replicate the time-course of development and learning in multisensory perception, thereby expanding the range of experimental data to which the model could be applied.

Notes

1. An anonymous reviewer noted that the four hypotheses are not necessarily logically distinct because Hypothesis 2 (training in multisensory environments leads to better learning than training in unisensory environments) is a special case of Hypothesis 1 (sensory integrations provides significant statistical advantages), and because Hypothesis 3 (neural representations of events are often modality invariant) is a special case of Hypothesis 4 (representations based on different modalities are associated with each other).
2. As a rule, this statement is correct. However, there are exceptions to this rule. That is, there exist parametric methods which infer the dimensionality of representations. The interested reader should see Green and Richardson (2001) and Rasmussen and Ghahramani (2001).
3. The Battacharyya distance is a statistical metric measuring the similarity of two discrete or continuous probability distributions. For discrete distributions p and q defined

over the domain X, it equals $-\ln\left(\sum_{x \in X} \sqrt{p(x)q(x)}\right)$. We also considered the Kullback-Leibler distance, but use of this metric led to numerical instabilities.

4. We followed the simulation procedures of Navarro and Griffiths (2008) with the exception that we used a uniform distribution over the interval [0, 1] as a prior distribution on the weights instead of a gamma distribution.

Acknowledgments

References

Abidi, M. A., & Gonzalez, R. C. (1992). *Data fusion in robotics and machine intelligence*. San Diego, CA: Academic Press.

Adams, R. P., Wallach, H., & Ghahramani, Z. (2010). Learning the structure of deep sparse graphical models. In Y. W. Teh & M. Titterington (Eds.), *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics* (pp. 1–8).

Alais, D., & Burr, D. (2004). The ventriloquist effect results from near-optimal bimodal integration. *Current Biology*, *14*, 257–262.

Alais, D., Newell, F. N., & Mamassian, P. (2010). Multisensory processing in review: From physiology to behavior. *Seeing and Perceiving*, *23*, 3–38.

Alvarado, J. C., Rowland, B. A., Stanford, T. R., & Stein, B. E. (2008). A neural network model of multisensory integration also accounts for unisensory integration in superior colliculus. *Brain Research*, *1242*, 13–23.

Amedi, A., Malach, R., Hendler, T., Peled, S., & Zohary, E. (2001). Visuo-haptic object-related activation in the ventral visual pathway. *Nature Neuroscience*, *4*, 324–330.

Anastasio, T. J., Patton, P. E., & Belkacem-Boussaid, K. (2000). Using Bayes' rule to model multisensory enhancement in the superior colliculus. *Neural Computation*, *12*, 1165–1187.

Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Erlbaum.

Austerweil, J. L., & Griffiths, T. L. (2009). Analyzing human feature learning as nonparametric Bayesian inference. In D. Koller, Y. Bengio, D. Schuurmans, & L. Bottou (Eds.), *Advances in neural information processing systems 21*. Cambridge, MA: MIT Press.

Barlow, H. B. (1961). Possible principles underlying the transformation of sensory messages. In W. Rosenblith (Ed.), *Sensory communication* (pp. 217–234). Cambridge, MA: MIT Press.

Battaglia, P. W., Jacobs, R. A., & Aslin, R. N. (2003). Bayesian integration of visual and auditory signals for spatial localization. *Journal of the Optical Society of America A*, *20*, 1391–1397.

Blei, D. M., Griffiths, T. L., & Jordan, M. I. (2010). The nested Chinese restaurant process and Bayesian inference of topic hierarchies. *Journal of the ACM*, *57*, 1–30.

Calvert, G. A., Bullmore, E. T., Brammer, M. J., Campbell, R., Williams, S. C. R., McGuire, P. K., Woodruff, P. W. R., Iversen, S. D., & David, A. S. (1997). Activation of auditory cortex during silent lipreading. *Science*, *276*, 593–596.

Chater, N., & Oaksford, M. (1999). Ten years of the rational analysis of cognition. *Trends in Cognitive Sciences*, *3*, 57–65.

Chater, N., & Oaksford, M. (2008). *The probabilistic mind*. Oxford, UK: Oxford University Press.

Clark, J. J., & Yuille, A. L. (1990). *Data fusion for sensory information processing systems*. Boston: Kluwer Academic Publishers.

Cooke, T., Jäkel, F., Wallraven, C., & Bülthoff, H. H. (2007). Multimodal similarity and categorization of novel, three-dimensional objects. *Neuropsychologia*, *45*, 484–495.

Courville, A., Eck, D., & Bengio, Y. (2009). An infinite factor model hierarchy via a noisy-or mechanism. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, & A. Culotta (Eds.), *Advances in neural information processing systems 22* (pp. 405-413).

Cover, T. M., & Thomas, J. A. (1991). *Elements of information theory*. New York: Wiley.

Cox, R. T. (1961). *The algebra of probable inference*. Baltimore: Johns Hopkins University Press.

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, *415*, 429–433.

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

Fetsch, C. R., DeAngelis, G. C., & Angelaki, D. E. (2010). Visual-vestibular cue integration for heading perception: Applications of optimal cue integration theory. *European Journal of Neuroscience*, *31*, 1721–1729.

Geisler, W. S. (2004). Ideal observer analysis. In L. M. Chalupa & J. S. Werner (Eds.), *The visual neurosciences* (pp. 825–837). Cambridge, MA: MIT Press.

de Gelder, B., & Vroomen, J. (2000). The perception of emotions by ear and by eye. *Cognition and Emotion*, *14*, 289–311.

de Gelder, B., & Vroomen, J. (2003). Multisensory integration, perception, and ecological validity. *Trends in Cognitive Sciences*, *7*, 460–466.

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

Gepshtein, S., & Banks, M. S. (2003). Viewing geometry determines how vision and touch combine in size perception. *Current Biology*, *13*, 483–488.

Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (1996). *Markov Chain Monte Carlo in practice*. London, UK: Chapman & Hall.

Goldstone, R. L. (1998). Perceptual learning. *Annual Review of Psychology*, *49*, 585–612.

Green, P., & Richardson, S. (2001). Modeling heterogeneity with and without the Dirichlet process. *Scandinavian Journal of Statistics*, *28*, 355–377.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

Griffiths, T. L., & Ghahramani, Z. (2005). Infinite latent feature models and the Indian buffet process. Gatsby Unit Technical Report GCNU-TR-2005-001.

Griffiths, T. L., & Ghahramani, Z. (2006). Infinite latent feature models and the Indian buffet process. In B. Schölkopf, J. Platt, & T. Hofmann (Eds.), *Advances in neural information processing systems 18*. Cambridge, MA: MIT Press.

Griffiths, T. L., Kemp, C., & Tenenbaum, J. B. (2008). Bayesian models of cognition. In R. Sun (Ed.), *The Cambridge handbook of computational psychology*. New York: Cambridge University Press.

Hershey, J., & Movellan, J. R. (1999). Audio-vision: Locating sounds via audio-visual synchrony. In *Proceedings of the 6th symposium on neural computation* (pp. 57–63). Los Angeles, CA: California Institute of Technology.

Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge, UK: Cambridge University Press.

Knill, D. C., & Richards, W. (1996). *Perception as Bayesian inference*. Cambridge, UK: Cambridge University Press.

Konkle, T., Wang, Q., Hayward, V., & Moore, C. I. (2009). Motion aftereffects transfer between touch and vision. *Current Biology*, *19*, 745–750.

von Kriegstein, K., & Giraud, A.-L. (2006). Implicit multisensory associations influence voice recognition. *PLoS Biology*, *4*, e326.

Lacey, S., Campbell, C., & Sathian, K. (2007). Vision and touch: Multiple or multisensory representations of objects? *Perception*, *36*, 1513–1521.

Lacey, S., Hall, J., & Sathian, K. (2010). Are surface properties integrated into visuohaptic object representations? *European Journal of Neuroscience*, *31*, 1882–1888.

Landy, M. S., Maloney, L. T., Johnston, E. B., & Young, M. J. (1995). Measurement and modeling of depth cue combination: In defense of weak fusion. *Vision Research*, *35*, 389–412.

Lehmann, S., & Murray, M. M. (2005). The role of multisensory memories in unisensory object discrimination. *Cognitive Brain Research*, *24*, 326–334.

Marr, D. (1982). *Vision*. New York: Freeman.

Massaro, D. W. (1998). *Perceiving talking faces: From speech perception to a behavioral principle*. Cambridge, MA: MIT Press.

Massaro, D. W., Cohen, M. M., Campbell, C. S., & Rodriguez, T. (2001). Bayes factor of model selection validates FLMP. *Psychonomic Bulletin & Review*, *8*, 1–17.

Movellan J. R. (1995). Visual speech recognition with stochastic networks. In G. Tesauro, D. Toruetzky, & T. Leen (Eds.), *Advances in neural information processing systems 7*. Cambridge, MA: MIT Press.

Navarro, D. J., & Griffiths, T. L. (2008). Latent features in similarity judgments: A nonparametric Bayesian approach. *Neural Computation*, *20*, 2597–2628.

Navarro, D. J., Griffiths, T. L., Steyvers, M., & Lee, M. D. (2006). Modeling individual differences using Dirichlet processes. *Journal of Mathematical Psychology*, *50*, 101–122.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, *9*, 249–265.

Neapolitan, R. E. (2004). *Learning Bayesian networks*. Upper Saddle River, NJ: Pearson Prentice Hall.

Oaksford, M., & Chater, N. (1999). *Rational models of cognition*. Oxford, UK: Oxford University Press.

Pekkola, J., Ojanen, V., Autti, T., Jääskeläinen, I. P., Möttönen, R., Tarkiainen, A., & Sams, M. (2005). Primary auditory cortex activation by visual speech: An fMRI study at 3T. *Neuroreport*, *16*, 125–128.

Pouget, A., Deneve, S., & Duhamel, J. R. (2002). A computational perspective on the neural basis of multisensory spatial representations. *Nature Review Neuroscience*, *3*, 741–747.

Quiroga, R. Q., Kraskov, A., Koch, C., & Fried, I. (2009). Explicit encoding of multimodal percepts by single neurons in the human brain. *Current Biology*, *19*, 1308–1313.

Rasmussen, C. E. (2000). The infinite Gaussian mixture model. In S. A. Solla, T. K. Leen, & Müller, K.-R. (Eds.), *Advances in neural information processing systems 12* (pp. 554–560). Cambridge, MA: MIT Press.

Rasmussen, C. E., & Ghahramani, Z. (2001). Occam's razor. In T. K. Leen, T. G. Dietterich, & V. Tresp (Eds.), *Advances in neural information processing systems 13* (pp. 294–300). Cambridge, MA: MIT Press.

Sathian, K., Zangaladze, A., Hoffman, J. M., & Grafton, S. T. (1997). Feeling with the mind's eye. *Neuroreport*, *8*, 3877–3881.

Seitz, A. R., Kim, R., & Shams, L. (2006). Sound facilitates visual learning. *Current Biology*, *16*, 1422–1427.

Shams, L., & Seitz, A. R. (2008). Benefits of multisensory learning. *Trends in Cognitive Sciences*, *12*, 411–417.

Shepard, R. N., & Arabie, P. (1979). Additive clustering: Representation of similarities as combinations of discrete overlapping properties. *Psychological Review*, *86*, 87–123.

Tanabe, H. C., Honda, M., & Sadato, N. (2005). Functionally segregated neural substrates for arbitrary audio-visual paired-association learning. *Journal of Neuroscience*, *25*, 6409–6418.

Todorov, E. (2004). Optimality principles in sensorimotor control. *Nature Neuroscience*, *7*, 907–915.

Wallace, M. T., & Stein, B. E. (2001). Sensory and multisensory responses in the newborn monkey superior colliculus. *Journal of Neuroscience*, *21*, 8886–8894.

Williamson, S., Orbanz, P., & Ghahramani, Z. (2010). Dependent India buffet processes. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics* (pp. 924–931).

Wozny, D. R., Beierholm, U. R., & Shams, L. (2008). Human trimodal perception follows optimal statistical inference. *Journal of Vision*, *8*(3), 1–11.

Zangaladze, A., Epstein, C. M., Grafton, S. T., & Sathian, K. (1999). Involvement of visual cortex in tactile discrimination of orientation. *Nature*, *401*, 587–590.

Zhou, Y. D., & Fuster, J. M. (2000) Visuo-tactile cross-modal associations in cortical somatosensory cells. *Proceedings of the National Academy of Sciences of the United States of America*, *97*, 9777–9782.