A Bayesian Nonparametric Approach to Multisensory Perception

İlker Yıldırım (iyildirim@bcs.rochester.edu)
Robert A. Jacobs (robbie@bcs.rochester.edu)
Department of Brain & Cognitive Sciences, University of Rochester, Rochester, NY 14627, USA

Abstract

We propose a Bayesian nonparametric model of multisensory perception based upon the Indian buffet process. The model includes a set of latent variables that learn multisensory features from unisensory data. The model is highly flexible because it makes few statistical assumptions. In particular, the number of latent multisensory features is not fixed a priori. Instead, this number is estimated from the observed data. We applied the model to a real-world visual-auditory data set obtained when people spoke English digits. Our results are consistent with several hypotheses about multisensory perception from the cognitive neuroscience literature. We found that the model obtained the statistical advantages provided by sensory integration. We also found that the model acquired multisensory representations that were relatively sensory invariant. Lastly, we found that the model was able to associate unisensory representations based on different modalities.

Keywords: multisensory perception; Bayesian modeling; rational analysis; Indian buffet process

Introduction

We learn about our environments from many different senses. Objects can be seen, heard, touched, tasted, and smelled. How are our mental representations based on these different sensory modalities structured, combined, and coordinated?

Cognitive neuroscientists have recently studied three important hypotheses about multisensory perception. First, researchers have conjectured that multisensory representations are advantageous because sensory integration ameliorates the effects of bias and 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 maximum likelihood 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 was obtained by Ernst and Banks (2002), for example, who found that people's estimates of object height based on both visual and haptic information was more reliable than their estimates based on either visual or haptic information alone.

Second, researchers have hypothesized that our neural representations of objects are often sensory invariant, meaning they are the same (or at least similar) regardless of the sensory modalities through which we perceive those objects. Evidence consistent with this hypothesis was obtained by Amedi et al. (2001). They showed that a neural region known as the lateral occipital complex (LOC) shows similar patterns of activation regardless of whether an object is seen or touched.

Third, researchers have speculated that representations based on different modalities are associated with each other. Suppose that an observer sees, but does not hear, an object. A visual representation of that object will be active in the observer's brain, and this representation will often predict or activate an auditory representation of the object even though the object is not heard. Evidence consistent with this hypothesis was obtained by Calvert et al. (1997). They found that viewing facial movements associated with speech (lipreading) leads to activation of auditory cortex in the absence of auditory speech sounds.

Here, we propose a model of multisensory perception that learns about its multisensory environment 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. Examples include factor analysis, where the number of latent variables is fixed a priori, and cluster analysis, where the number of clusters is fixed a priori. In contrast, a nonparametric method uses representations that are allowed to grow structurally as more data are observed. These methods are often used when the goal is to impose as few assumptions as possible and to "let the data speak for themselves" (Blei, Griffiths, & Jordan, 2010). Examples include Dirichlet process mixture models (or Chinese restaurant processes) and Indian buffet processes.

The proposed model of multisensory perception is an instance of a Bayesian nonparametric model. It "explains" the unisensory representations arising from different modalities through the use of a set of latent or hidden variables that learn multisensory representations. 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 representations are estimated from the observed unisensory data, nonparametric statistical methods are required for inference. We use a Bayesian nonparametric framework developed by Griffiths and Ghahramani (2005, 2006) known as the Indian buffet process. Due to its Bayesian foundations, the proposed model can be regarded as an ideal observer model inferring optimal features of its multisensory environment (Austerweil & Griffiths, 2009).

We applied the proposed model to a visual-auditory data set obtained when people spoke different digits. Our results are consistent with the three hypotheses from the cognitive neuroscience literature described above. It was found that the model obtained the statistical advantages provided by sensory integration: categorization of objects was more accurate based on its latent multisensory representations than on the latent features of unisensory models. In addition, the model's latent or multisensory representations were relatively sensory invariant. That is, similar representations of an object were formed regardless of whether an object was seen or heard. Lastly, the model was able to associate representations based on different modalities. In other words, it could use one type of unisensory representation to predict or activate another type of unisensory representation.

Visual-Auditory Data Set

The multisensory perception model was applied to a visual-auditory data set known as the Tulips1 data set (Movellan, 1995). Twelve people (9 adult males, 3 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 6 visual features: the width and height of the outer corners of the mouth, the width and height of the 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[1], 1 log-power, 12 cepstral coefficient derivatives, and 1 log-power derivative. Because speech utterances had different durations, we sampled 6 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 $\times$ 6 visual features per frame) and 156 auditory features (6 frames $\times$ 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.

Multisensory Perception Model

We describe the proposed model in the context of a visual-auditory environment, though we note that the model is equally applicable to other sensory modalities and to any number of modalities. A coarse schematic of the model is illustrated in Figure 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 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.

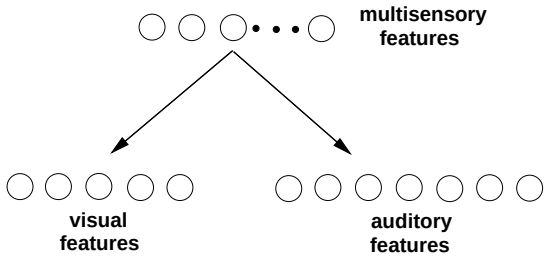

Figure 1: A coarse schematic of the multisensory perception model.

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 in Figure 2. 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:

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

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:

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

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 N^{th} harmonic

[1] Cepstral coefficients are the coefficients of the Fourier transform representation of the log magnitude spectrum.

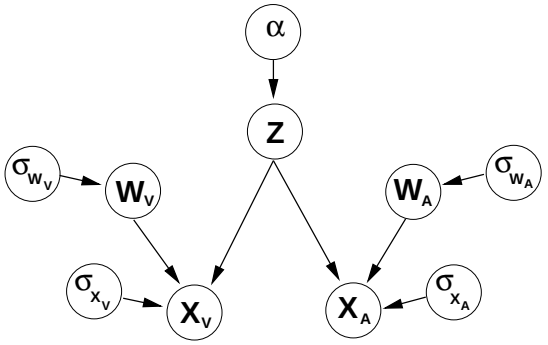

Figure 2: A Bayesian network representation of the multisensory perception model.

number, m_k is the number of objects with feature k, and α is a variable influencing the number of features.

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, $z_i W_\beta$, 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:

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

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

Simulation Results

The multisensory perception model was applied to the visual-auditory data set. To better understand its performances, 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 visual-auditory data set, it received only the visual features. Similarly, the auditory-only model contains only two sets of variables corresponding to auditory and latent features. It received only the auditory features from the data set.

Because exact inference in the models is computationally intractable, approximate inference using Markov chain Monte Carlo (MCMC) sampling methods (e.g., Gelman et al., 1995) was performed based upon the training data following Griffiths and Ghahramani (2005). A single chain of each model was simulated. Each chain was run for 5000 iterations. The first 3000 iterations were discarded as burn-in. To reduce correlations among variables at nearby iterations, the remaining iterations were thinned to every 10^{th} iteration (i.e., only variable values at every 10^{th} iteration were retained). Thus, the results below are based on 200 iterations.

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 Figure 3 show the distributions of the numbers of latent features in the visual-only, auditory-only, and multisensory models. The visual-only model used relatively few latent features, the auditory-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.

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 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 $\rightarrow$ digit 3, cluster B $\rightarrow$ 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 Figure 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 auditory-only model showed better performance, and the multisensory model showed the best performance.

Its 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 visual-only or auditory-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 Figure 4. The mixed model showed significantly poorer performance than the multisensory model, thus suggesting the statistical advantages of segregating perceptual inputs into separate streams.

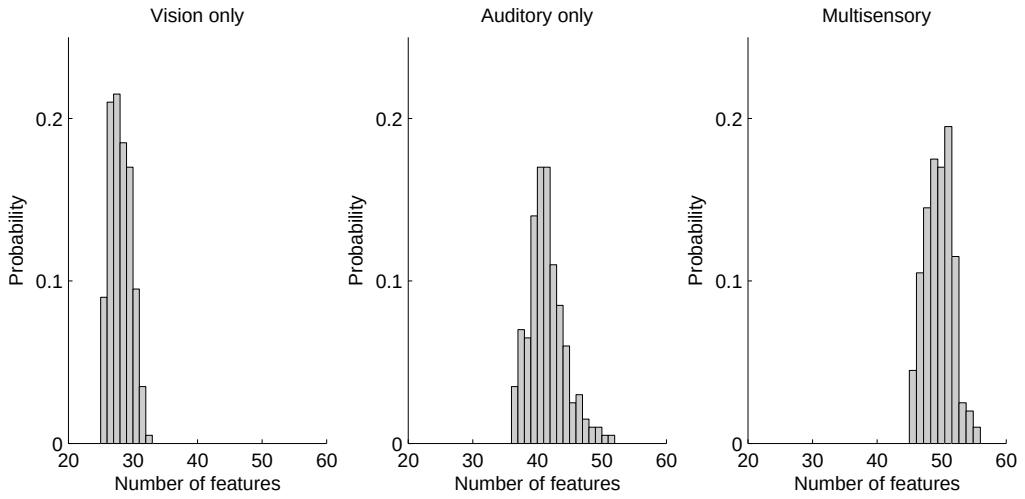

Figure 3: The distributions of the numbers of latent features in the visual-only (left), auditory-only (middle), and multisensory (right) perception models.

This analysis was repeated using the data items in the test set. Performing the analysis 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 our MCMC sampler on a 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. For a given model, 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 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 an exhaustive 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 Figure 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.

Sensory invariance

As discussed above, neural representations of objects are often sensory invariant. 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. 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. Recall that these are the latent or multisensory representations with non-zero 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 $\mathcal{L}_i$ 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 between $p(Z|X_V)$ and $p(Z|X_A)$.[2] 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 $\mathcal{L}_i$ 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 25$^{\text{th}}$, 50$^{\text{th}}$, and 75$^{\text{th}}$ percentiles, respectively. By way of comparison, we also computed the distance between $p(Z|X_A)$ and a uniform distribution over multisensory representations. The distribu-

[2] We also considered the Kullback-Leibler distance but use of this metric led to numerical instabilities.

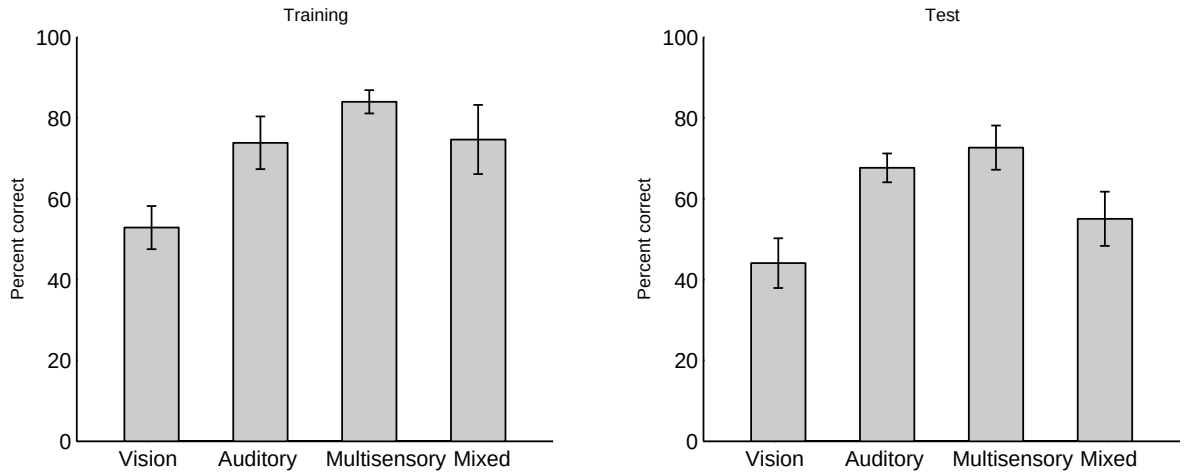

Figure 4: Categorization performances of the vision-only, auditory-only, multisensory, and mixed models on the training set (left) and on the test set (right). The horizontal axis of each graph gives the model, and the vertical axis plots the percent of data items correctly classified (error bars indicate the standard deviations of these percents across iterations of an MCMC chain).

tion of these distances has values of 3.49, 7.83, and 19.04 as its 25^{th}, 50^{th}, and 75^{th} percentiles.

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

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). 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 show this behavior?

This question was studied using the data items in the test set. Let $\mathcal{V}$ and $\mathcal{A}$ denote the sets of visual and auditory feature representations for the data items in the training set. Once again, let $\mathcal{L}_i$ 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 $\mathcal{A}$ and $\mathcal{L}_i$. Analogous computations were carried out to compute the distibution of a visual representation given an item's auditory features.

Representative results are shown in Figure 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.

Conclusions

Bayesian nonparametric approaches to modeling are becoming increasingly popular in the cognitive science and machine learning literatures. We regard this approach as an important advance over conventional parametric approaches in which a researcher sets the number of latent variables by hand, often in an ad hoc or unprincipled manner. How can a researcher be sure that the number of latent features should, for example, be exactly 10? Shouldn't the number of latent features be determined by the structure of the task or data set? The Bayesian nonparametric approach is also an advance over modeling approaches that define a set of models, each with a different number of latent features, and perform "model comparison" to select the best model. 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.

We have proposed a Bayesian nonparametric model of multisensory perception. The model includes a set of latent vari-

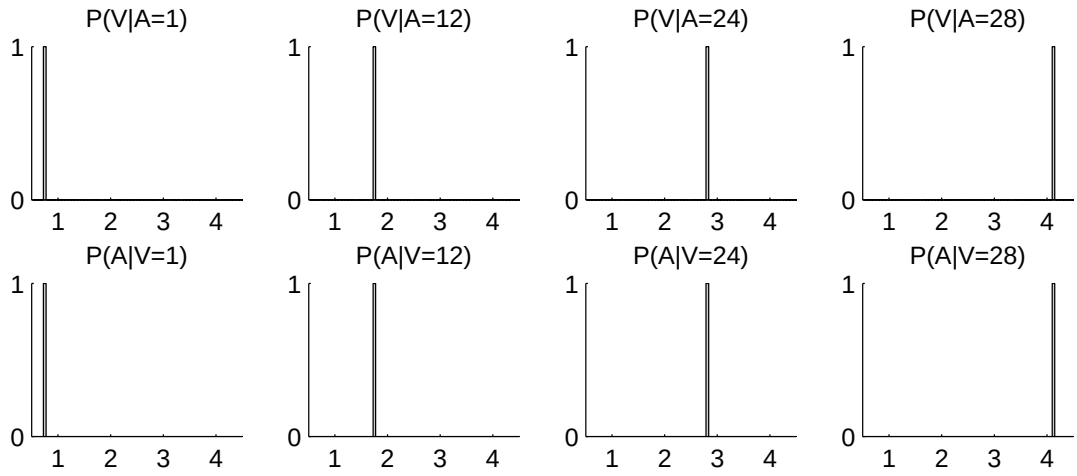

Figure 5: Graphs in the top row demonstrate that when presented with auditory features of a test item corresponding to one of the digits, the multisensory perception model's distribution of visual representations was tightly peaked at a representation corresponding to the same digit. Graphs in the bottom row show analogous results for distributions of auditory representations given test items' visual features.

ables that learn multisensory features from unisensory data. The model is highly flexible because it makes few statistical assumptions. In particular, the number of multisensory features is not fixed a priori. Instead, this number is estimated from the data.

We applied the model to a real-world visual-auditory data set obtained when people spoke English digits. Our results are consistent with several hypotheses about multisensory perception from the cognitive neuroscience literature. We found that the model obtained the statistical advantages provided by sensory integration. We also found that the model acquired multisensory representations that were relatively sensory invariant. Lastly, we found that the model was able to associate unisensory representations based on different modalities.

Because the multisensory perception model is based on Bayesian statistics, it can be regarded as an ideal observer inferring optimal multisensory features from unisensory data (Austerweil & Griffiths, 2009). As such, it provides a basis for a rational analysis of multisensory perception. This analysis suggests that the acquisition of latent multisensory representations that are sensory invariant and more statistically robust than latent features from unisensory models is a rational response of an agent attempting to learn the structure of its multisensory environment. It also suggests the rationality of acquiring associations among unisensory representations.

Acknowledgments

We thank J. Drugowitsch, A. E. Orhan, and C. Sims for many helpful discussions, and J. Movellan for making the Tulips1 data set available on the web. This work was supported by a research grant from the National Science Foundation (DRL-0817250).

References

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.

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.

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.

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

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

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.

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.