available at www.sciencedirect.com

ScienceDirect

www.elsevier.com/locate/brainres

BRAIN RESEARCH

Review

Linking neurons to behavior in multisensory perception: A computational review

Wei Ji Ma, Alexandre Pouget*

Department of Brain and Cognitive Sciences, University of Rochester, Rochester NY 14627, USA

ARTICLE INFO

Article history:
Accepted 27 April 2008
Available online 21 June 2008

Keywords:
Multisensory integration
Cue combination
Computational modeling
Population coding
Neural variability
Superadditivity
Segregation

ABSTRACT

A large body of psychophysical and physiological findings has characterized how information is integrated across multiple senses. This work has focused on two major issues: how do we integrate information, and when do we integrate, i.e., how do we decide if two signals come from the same source or different sources. Recent studies suggest that humans and animals use Bayesian strategies to solve both problems. With regard to how to integrate, computational studies have also started to shed light on the neural basis of this Bayes-optimal computation, suggesting that, if neuronal variability is Poisson-like, a simple linear combination of population activity is all that is required for optimality. We review both sets of developments, which together lay out a path towards a complete neural theory of multisensory perception.

Contents

Accurate perception frequently relies on combining uncertain information from multiple senses. Imagine that you are trying to locate a predator hiding in the bushes. You hear a faint sound of the predator's footsteps and at the same time you see a movement of the leaves. That movement could be caused by the animal, but also by a gust of wind. If the predator caused the movement, the visual information will help you localize it with greater precision.

This example illustrates several general aspects of multisensory perception. Combining information across senses can

* *Corresponding author*. Department of Brain and Cognitive Sciences University of Rochester Rochester, NY 14627, USA. Fax: +1 585 442 9216.
E-mail address: weijima@gmail.com (W.J. Ma).

0006-8993/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.brainres.2008.04.082

be of critical importance to an animal's survival, making it plausible that evolutionary pressure has optimized the neural circuits that serve this purpose. Moreover, those circuits have to solve two problems simultaneously: figuring out whether two cues had the same source (the predator) or different sources (the predator and the wind), and in the former case, how to combine them. Finally, cues can come with different reliabilities. Visual information will be more reliable on a sunny day than on a foggy day, and you can trust auditory information more if there is little background noise. These aspects have guided the theoretical developments we discuss in this review.

1. Optimal cue integration

When a common source is assumed, a systematic strategy to quantify cue combination is to introduce a small discrepancy (also called conflict, disparity, or incongruency) between the cues. The conflict must be small in order to not violate the common-source assumption. In such a paradigm, the percept (estimate of the stimulus) inferred from both cues presented together will lie somewhere in between the percepts inferred from each cue individually. The intuition is that higher weight will be given to the most reliable cue, and that therefore the multi-cue percept will be closest to the percept obtained from that cue. Recent psychophysical studies have quantified this intuition, both across (Alais and Burr, 2004; Battaglia et al., 2003; Ernst and Banks, 2002; van Beers et al., 1996; Wolpert et al., 1995) and within sensory modalities (Jacobs, 1999; Knill and Saunders, 2003). As an example, we consider a laboratory version of the ventriloquist effect (Alais and Burr, 2004), the well-known illusion in which a performer makes a puppet appear to speak (Howard and Templeton, 1966; Welch and Warren, 1980). This experiment involved spatial localization along the azimuthal dimension, based on brief visual flashes and auditory clicks. Importantly, observers were instructed to regard each pair of multisensory signals as being caused by a single, well-localized event, for instance a ball hitting the screen. The investigators found that the mean auditory–visual estimates of location, locations $\hat{s}_{AV}$, could be expressed as a linear combination of the auditory and visual s_A and s_V:

$$\hat{s}_{AV} = \frac{w_A s_A + w_V s_V}{w_A + w_V} \tag{1}$$

In this expression, the weights are given by the inverse variances of estimates in the respective modalities: $w_A = \frac{1}{\sigma_A^2}$ and $w_V = \frac{1}{\sigma_V^2}$. For example, if in a certain condition the visual variance is larger than the auditory variance (and therefore vision is less reliable than audition), vision will be given less weight than audition in the combination.

Moreover, the inverse variance of the auditory–visual estimates was found to be

$$\frac{1}{\sigma_{AV}^2} = \frac{1}{\sigma_A^2} + \frac{1}{\sigma_V^2} \tag{2}$$

This indicates that using two cues led to higher precision than using any one cue. The right-hand side of Eq. (2) also gives the highest possible precision that can be achieved by an unbiased estimator, given σ_A and σ_V. Eqs. (1) and (2) state what is meant by statistical optimality in this task. Although they summarize average human behavior over many trials (they give the mean and variance of maximum-likelihood estimates), it is commonly assumed that they reflect regularities that hold on a trial-by-trial basis. On a single trial, we can think of a sensory cue as providing a probability distribution over the stimulus. If we denote the auditory-only distribution by $p(s \mid A)$, the visual-only distribution by $p(s|V)$ and the multisensory distribution by $p(s|A, V)$, then the optimal multisensory distribution is the product distribution (Clark and Yuille, 1990; Yuille and Bulthoff, 1996)

$$p(s|A,V) \propto p(s|A)p(s|V), \tag{3}$$

where the proportionality is such that $p(s \mid A,V)$ is normalized to 1. We have assumed that the auditory and visual distributions are independent given the stimulus (this is called conditional independence). When the distributions in Eq. (3) are Gaussian, Eqs. (1) and (2) directly follow from Eq. (3). As human behavior follows Eqs. (1) and (2) in a wide variety of paradigms, (multisensory) cue integration has become a poster child of Bayes-optimal computation.

Several years ago, a review article stated that these findings of approximate Bayes-optimal cue integration in humans raised two central questions (Banks, 2004): "1. how does the brain know the variances of its sensory estimates to make the correct weight assignments; 2. how does the brain know when sensory estimates are coming from the same source and not different sources, so that combining makes sense?" Since then, significant progress has been made on both these questions, in particular in the theoretical domain.

2. Optimal cue integration with neural populations

When studying how neuronal circuits implement near-optimal cue integration, an important fact to take into account is that the responses of cortical neurons are typically very variable (Compte et al., 2003; Dean, 1981; Holt et al., 1996; Tolhurst et al., 1982). Presenting the same stimulus repeatedly will give rise to many different population responses. A first sight, such variability is a nuisance that could compromise optimality. Recent work, however, has argued that the presence of variability is not the problem. If we experience uncertainty about a stimulus, this stimulus must generate variability in the brain, otherwise there would be no uncertainty. However, the format of the neural variability is important in the neural implementation of the optimal cue integration (Eq. (3)) (Ma et al., 2006). If the statistics of the variability are known (either to the experimenter or to downstream neurons), then Bayes' rule can be used to convert the population pattern of activity on a single trial into a probability distribution over the stimulus. To be precise, if the population activity on a single trial is denoted by a vector $\mathbf{r}=(r_1, r_2, \ldots, r_N)$, where r_i is the activity of the ith neuron and N is the number of neurons, then one can obtain the so-called posterior distribution through

$$p(s|r) \propto p(\mathbf{r}|s)p(s), \tag{4}$$

where $p(\mathbf{r}|s)$ is the response distribution and $p(s)$ is the prior distribution (Foldiak, 1993; Sanger, 1996). The posterior

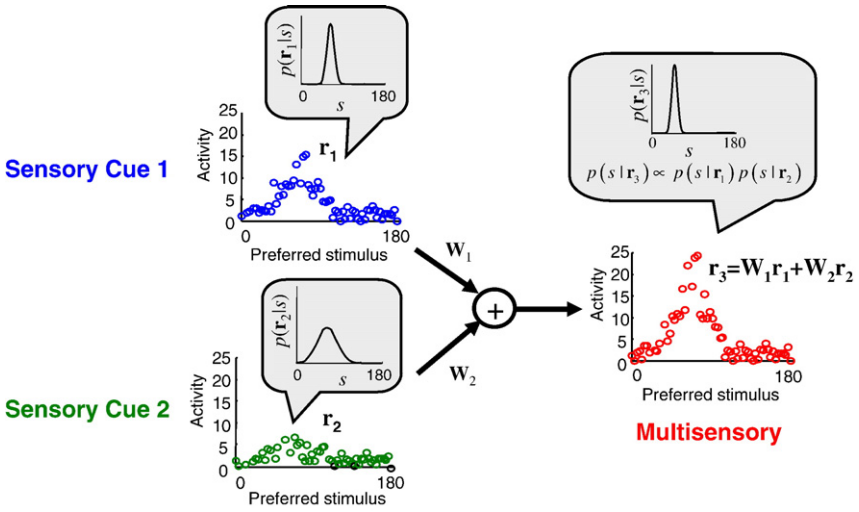

Fig. 1 – Schematic of a network that can perform optimal multisensory cue integration on each trial without having to estimate the reliabilities of cues. A simple linear combination of the population patterns of activity elicited by the cues guarantees optimality, as long as neuronal variability is *Poisson-like* (see text). This allows for correlated, non-Poisson variability. The dialogue boxes show the probability distributions over the stimulus that are encoded in each population on a single trial. The synaptic weights W_1 and W_2 have to be chosen according to the tuning curves and covariance matrices in the input layers, but can then stay fixed across trials.

distribution does not only reflect the most probable value of the stimulus (the maximum-a-posteriori estimate), but also the observer's uncertainty (through the width of the distribution). This form of neural representation is called a probabilistic population code.

In probabilistic population codes, the uncertainty about s is related to the variability in $\mathbf{r}$. It becomes therefore essential that we characterize precisely the distribution $p(\mathbf{r} \mid s)$. Given the available neural data (Gur et al., 1997; Tolhurst et al., 1982), it appears that $p(\mathbf{r}|s)$ is well approximated by the exponential family with linear sufficient statistics, a family we call 'Poisson-like' for short, and which includes independent Poisson variability as a special case. Surprisingly, when neural variability is in this Poisson-like family, it can be shown that optimal cue integration – which is a multiplicative operation at the level of the posteriors, see Eq. (3) – is realized through a simple linear combination of population responses (Ma et al., 2006), see Fig. 1. This holds irrespective of the shapes of the tuning curves, or the covariance matrices. It was found that such a linear combination can also be implemented in a biophysically realistic neural network of conductance-based integrate-and-fire neurons. The fact that this neural operation does not require the estimation of variance at any point provides an answer to the first question stated above. The variances are automatically taken into account appropriately through the interplay of neural variability (Poisson-like) and network operations (linear combination). The format of the neural variability is key in this framework: it facilitates optimal computation. This theory still requires downstream neurons to collapse the distribution onto a single value when an action is required, such as a saccadic eye movement. Again under the assumption of Poisson-like variability, this read-out can be done optimally using a line attractor network (Deneve et al., 1999; Latham et al., 2003). However, importantly, all information about uncertainty is preserved until this very last stage in sensorimotor processing.

3. Multisensory integration in relation to other probabilistic computations

The appeal of probabilistic population codes, in which a population pattern of activity encodes the certainty about a stimulus, is that they are not limited to multisensory perception. Ecologically important tasks often require combining pieces of uncertain sensory information with each other or with prior information. In multisensory perception, cues from different modalities get combined. Examples in other domains include perceptual decision-making (combining information over time and selecting an action), visual search (detecting or localizing a target by integrating information over space), visual working memory (for example, over space and time to detect a change), and sensorimotor control (combining sensory information with an internal model). In each of these domains, the brain needs to take into account the certainty about various pieces of information in order to perform the task optimally. Cue integration offers merely the simplest illustration of this: if the reliabilities of the cues were not encoded, optimal cue combination would be impossible. As a consequence, it is often insufficient if a population of neurons only encodes a single number, an estimate of the stimulus. Instead, it is necessary to represent certainty, or even better, an entire probability distribution over the stimulus, as in Eq. (4). Neural processing from layer to layer can then be constructed such that it implements optimal computations on those probability distributions, without ever making those distributions "explicit". Probabilistic population coding theory thus allows to ask in a very precise way what

neural operations correspond to given probabilistic operations at a behavioral level.

It remains to be seen to what extent humans exhibit near-optimal behavior in the above-mentioned paradigms, which require probabilistic computation. However, even where significant deviations from optimality exist, probabilistic population codes provide a constructive approach, as they enable a quantitative study of the sources of these deviations at the neural and behavioral levels in parallel. Probabilistic population coding theory thus places multisensory integration within a broad and encompassing framework together with other probabilistic computations.

4. Comparison with physiology

Using this theoretical framework, it is now possible to link optimal behavior to neural population activity. Imagine recording with a multi-electrode array from a population of multisensory neurons in an awake, behaving animal engaged in optimal cue combination (as tested behaviorally). Then the theory predicts that the response of multisensory neurons when two cues are presented is equal to the sum of their responses when each cue is presented separately (this follows from the fact that the multisensory output activity is a linear combination of the unisensory input activities). A problem with this simple prediction is that neurons may saturate as they approach their maximum firing rate. This saturation can be prevented by using global recurrent inhibition, that is, by subtracting a term proportional to the total activity. It can be shown that such inhibition does not affect optimality (Beck, Ma, et al, submitted). In other words, this scheme predicts

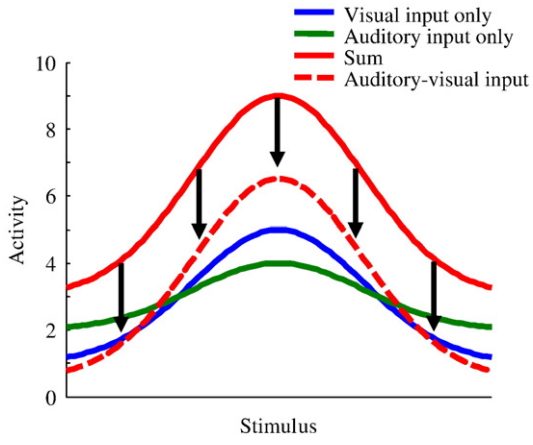

Fig. 2 – Prediction from the neural Bayesian model for the mean activity of a multisensory neuron as a function of the stimulus (arbitrary units) during optimal cue integration. The naïve prediction is that the response to multisensory input (solid red line) is the sum of the responses to unisensory inputs (blue and green lines). However, to keep neurons in their dynamic range, an arbitrary baseline may be subtracted (black arrows). This does not affect optimality. Consequently, neurons involved in optimal cue integration are expected to be additive (solid red line) or subadditive (dashed red line).

additivity or subaddivity (i.e., additivity minus a baseline shift) of the multisensory response, see Fig. 2. Moreover, the theory predicts that the information encoded in the multisensory population – as can be estimated using various decoding techniques (Averbeck et al., 2006; Nirenberg and Latham, 2003; Series et al., 2004) – should correlate with both the reliabilities of the unisensory inputs and the precision displayed behaviorally.

While such a conclusive experiment has not yet been performed, there is an abundance of recordings from single neurons in the superior colliculus that respond to both auditory and visual stimuli. Early reports of these neurons emphasized their superadditivity (meaning that the multisensory response exceeds the sum of the unisensory responses) (Meredith and Stein, 1986; Stein et al., 1988; Wallace et al., 1996) and this has since been a guiding concept in the field, even in behavioral and functional imaging studies (for reviews and criticism of its use in fMRI studies, see Beauchamp, 2005; Laurienti et al., 2005). In physiology, superadditivity is often invoked as evidence that a particular neuron is involved in multisensory integration. The neural Bayesian theory we have just outlined suggests an alternative: neurons involved in optimal stimulus inference are expected to display additive or subadditive responses. Interestingly, recent physiological work has shown that superadditive neurons constitute a minority and account for a relatively small number of spikes (Perrault et al., 2005; Populin and Yin, 2002; Stanford et al., 2005; Stanford and Stein, 2007). In contrast, the majority of multisensory neurons in the cat superior colliculus exhibit additivity or subadditivity, as predicted by the neural Bayesian theory.

This is not to say that superadditive neurons do not play a role in multisensory integration. They may be involved in other aspects of multisensory integration; for instance, it has been shown that nonlinear activation functions are needed for computations such as coordinate transformations and efficient read-out (Deneve et al., 1999; Deneve et al., 2001) and are found in multisensory neurons likely involved in those computations (Green and Angelaki, 2007). Interestingly, the nonlinear activation functions that are typically used in models (such as in Deneve et al., 1999; Deneve et al., 2001) predict the greatest nonlinearity for neurons whose unisensory response is weak, consistent with physiology (Stanford et al., 2005). It is also possible that superadditive neurons do not fire in the Poisson-like family, in which case optimal combination could involve a nonlinearity. Finally, the neural Bayesian theory outlined above does not deal with stimulus detection; it assumes that the stimulus is present and then extracts a posterior distribution over stimulus attributes (like position). In principle, we can extend our framework to detection, and it is quite possible that a nonlinearity will be required for optimality in this case. Multi-electrode recordings in awake, behaving animals should provide invaluable data to explore these issues further and to help test Bayesian theories that link physiology and behavior. The motivation for recording multiple neurons at once is that correlations can be estimated, which is essential to accurately estimate the information content of the population (Averbeck et al., 2006; Series et al., 2004).

Multi-electrode recordings could also be used to test other predictions from probabilistic population coding theory. For

instance, if neurons represent probability distributions as we have described, all information contained in the population should be recoverable with local linear decoders. In other words, both downstream neurons and experimentalists should find that nonlinear decoders do not perform any better than linear ones. This would not only make computation and learning in downstream layers considerably simpler, but also neural decoding by experimentalists, since nonlinear decoders require huge amounts of data to be tuned. In addition, this framework also predicts that one should be able to estimate, on a trial-by-trial basis, the confidence of the animal. For instance, if the posterior distribution extracted from the superior colliculus on a given trial predicts a precision of, say, 10° in saccadic endpoints, the animal should exhibit an average precision of about 10° on a large number of trials of the same type. To our knowledge, this prediction is specific to this framework.

Alternative theories have been put forward to explain superadditivity in these multisensory neurons in the superior colliculus, such as the one by Anastasio et al. (Anastasio et al., 2000; Patton and Anastasio, 2003). In this theory, the response of a single neuron is hypothesized to be proportional to the posterior probability of the stimulus being present in its receptive field. This model is somewhat limited in the type of data it can capture. For instance, the model assumes that neurons respond with the same firing rate whenever a stimulus is present, regardless of factors like position, contrast (for a visual target), or frequency (for auditory stimuli). Unfortunately, this is quite implausible. Moreover, this theory cannot explain why superior colliculus neurons fire with near-Poisson statistics: it assumes that incoming spike trains are Poisson, but predicts that output spike counts follow a very unusual distribution given target presence (or absence). As a result, it is difficult to assess whether this model truly accounts for the firing patterns of neurons in the superior colliculus, but further work might help to resolve this issue.

5. Cue combination without forced integration

The second question concerns the number of sources, or multiplicity, of multisensory cues. When an auditory and a visual stimulus are observed, they could have either the same source or different sources. In cue conflict experiments, the disparity between the cues is usually kept small, so that the observer has no difficulty imagining that they originate from the same source (forced integration). However, in natural circumstances, large disparities in space, time, or feature space occur frequently. In these conditions, observers will tend to perceive the signals as originating from different sources. This should be accounted for by a full theory of multisensory perception. Older and recent experiments have systematically investigated multisensory combination of simple stimuli in the presence of larger disparities, in humans (Bresciani et al., 2006; Choe et al., 1975; Kording et al., 2007; Roach et al., 2006; Slutsky and Recanzone, 2001; Thurlow and Jack, 1973; Wallace et al., 2004) and in cats (Rowland et al., 2007a). In these studies, observers performed one of four tasks: reporting the perceived origin of one stimulus while ignoring the other (Bresciani et al., 2005; Bresciani et al., 2006; Roach et al., 2006; Rowland et al., 2007a), reporting the perceived origins of the signals in both modalities separately (Kording et al., 2007; Shams et al., 2005), reporting whether they perceived the signals to originate from the same source (Choe et al., 1975; Jack and Thurlow, 1973; Slutsky and Recanzone, 2001), or reporting both the perceived origin of one stimulus and the perception of a common source (Wallace et al., 2004). Across this diversity of paradigms, it was found that the smaller the disparity, the greater the relative influence of the "irrelevant" modality (bias) and the greater the probability of perceiving a unified percept. Moreover, in the last paradigm, localization variability was smaller when observers reported a common source than when they reported different sources, and, paradoxically, bias was mostly negative when they reported different sources (Wallace et al., 2004).

These findings have been explained by a Bayesian causal inference model, proposed recently by two independent groups ((Beierholm et al., 2007; Kording and Tenenbaum, 2006; Kording et al., 2007; Sato et al., 2007); see also related models (Bresciani et al., 2006; Roach et al., 2006; Rowland et al., 2007a; Shams et al., 2005)). According to this model, an observer entertains two possible hypotheses about the process that generated the multisensory signals: that they have a common cause ($C=1$) or that they have separate, independent causes ($C=2$), see Fig. 3. On each trial, the observer computes the probability of each hypothesis, $p(C|x_A, x_V)$, based on the noisy sensory signals on that trial (x_A and x_V for auditory and visual, respectively) as well as prior information, $p(C)$, about the presence of a common cause:

$$p(C|x_A, x_V) \propto p(x_A, x_V|C)p(C) \quad (5)$$

In this equation, $p(x_A, x_V|C)$ is called the likelihood function of C. Subsequently, these probabilities (for $C=1$ and $C=2$) are used to weigh the stimulus estimates following from each

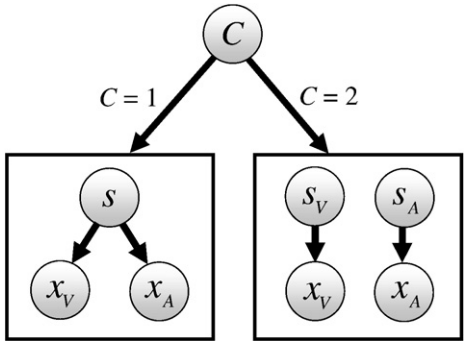

Fig. 3 – Generative model of two auditory–visual signals. The random variable C indicates the number of causes. If C=1, then the visual and auditory cues x_V and x_A have the same source – this is the generative model of traditional cue integration. If C=2, the stimuli have independent sources. On each trial, only x_V and x_A are available to the observer. The probability of each value of C is evaluated based on these cues and on prior information. This probability is then used to weigh the two hypotheses to optimally infer the source(s).

hypothesis (this follows from the Bayesian formalism). For example, the estimated location of the auditory stimulus, $\hat{s}_A$ is obtained from the estimate under the common-cause hypothesis, $\hat{s}_{common}$, and the estimate under the separate-causes hypothesis, $\hat{s}_{auditory}$,through (Kording et al., 2007)

$$\hat{s}_A = p(C=1 \mid x_A, x_V)\hat{s}_{common}(x_A, x_V) + p(C=2|x_A, x_V)\hat{s}_{auditory}(x_A) \quad (6)$$

Because the weights p(C=1|x_A,X_V) and p(C=2|x_A,x_V) depend in a nonlinear fashion on the cues, this scheme predicts nonlinear cue combination, even though $\hat{s}_{common}$ and $\hat{s}_{auditory}$ themselves are linear functions of x_A and $^{x}_{V}$ when all probability distributions over the stimulus are Gaussian. This model can quantitatively account for the auditory localization bias, unity judgments, and localization variability as a function of spatial disparity, as well as the paradoxical negative-bias effect (Beierholm et al., 2007; Kording and Tenenbaum, 2006; Kording et al., 2007; Sato et al., 2007) and the ventriloquism aftereffect (Sato et al., 2007). As such, it provides, at the behavioral level, a rigorous formulation of the heuristic known as the 'spatial principle' (Stein and Meredith, 1993), which states that multisensory stimuli are more likely or more effectively integrated when they originate from approximately the same spatial location.

Central to the causal inference model is the fact that the presence of a common source is not assumed as in traditional cue integration, but assigned a probability based on both single-trial evidence and prior information. Prior information can be manipulated by informing subjects of the existence of a disparity between the stimuli (Warren, 1979; Welch, 1972), which, as the model predicts, leads to a decrease in auditory localization bias. Furthermore, it has been argued that pairs of multisensory stimuli that tend to co-occur in natural environments should be treated differently from "arbitrary" pairs generated in the laboratory (De Gelder and Bertelson, 2003). In the causal inference model, this distinction can be captured by the prior probability of a common cause; all other things being equal, naturalistic pairs are expected to yield higher weights for the common-cause hypothesis. The model can be restated in the form of a mixed prior over pairs of stimuli (Beierholm et al., 2007; Kording et al., 2007) and is therefore an example of a mixture model (Yuille and Bulthoff, 1996), used earlier to describe the combination of depth cues (Knill, 2003). The causal inference model was compared to other Bayesian models and was found to describe the data best (Beierholm et al., 2007; Kording et al., 2007), although another model fared relatively well too (Roach et al., 2006). More work is needed to compare Bayesian models in other experimental paradigms.

Whereas the Bayesian causal inference model has so far been used to explain multisensory spatial localization, with spatial proximity being the cue for the existence of a common cause, it can easily be applied and extended to other feature judgments as well as other unity cues. Experiments show that increasing temporal disparity strongly decreases localization bias (Bertelson and Aschersleben, 1998; Radeau and Bertelson, 1987; Thomas, 1941), spatial and temporal disparities interact in eliciting the ventriloquism effect (Slutsky and Recanzone, 2001), and an analog to the ventriloquism effect occurs in the temporal dimension when stimuli are spatially coincident (Bertelson and Aschersleben, 2003), although spatial disparity fails to affect the latter effect (Vroomen and Keetels, 2006). In the Bayesian causal inference model, any feature used to infer whether two stimuli have the same origin will affect the likelihood of a common cause, and thus determine the relative weights of the hypotheses. This suggests that also the behavioral version of the so-called "temporal principle", stating that multisensory stimuli are more likely or more effectively integrated when they occur approximately simultaneously (Stein and Meredith, 1993), can be formalized in Bayesian terms.

The multiplicity aspect of multisensory perception (whether there are one or multiple sources) can be regarded as a problem of perceptual grouping (Bertelson, 1999; Radeau and Bertelson, 1987), i.e., deciding which elements of sensory input belong together and should be bound into a unified percept. It is similar to simultaneously segregating and identifying multiple sources within the same modality, a problem encountered in auditory scene analysis (Bregman, 1990; Feng and Ratnam, 2000) and in viewing superimposed patterns of moving dots with differing motion directions (Treue et al., 2000). A related case of perceptual grouping, contour integration, has been successfully described using Bayesian models (Elder and Goldberg, 2002; Feldman, 2001; Geisler et al., 2001). It is widely believed that Bayesian models of perceptual grouping hold promise as quantitative, probabilistic versions of Gestalt laws (Mamassian, 2006).

The neural implementation of Bayes-optimal causal inference in multisensory perception is as of yet unknown, but can be expected to be a generalization of the neural implementation of optimal cue integration (with forced fusion), i.e., linear combination for optimal posterior computation (Ma et al., 2006) and attractor dynamics for efficient estimation (Deneve et al., 1999; 2001). Doubly distributional population codes (Sahani and Dayan, 2003) have been proposed to address the problem of simultaneously encoding uncertain cues about multiple sources in the same modality in neural populations, and may be of use in causal inference as well.

Table 1 – Subproblems of multisensory integration from a computational view

Subproblem	Behavioral model	Neural theory
Taking into account reliabilities of cues	Bayes-optimal cue integration	Linear combination of population activities (assuming Poisson-like variability)
Efficient read-out	Maximum-likelihood estimation	Attractor dynamics (assuming Poisson-like variability)
One or multiple sources	Bayes-optimal causal inference	Unknown
Different coordinate frames	Coordinate transformation	Attractor dynamics, basis function network (assuming Poisson-like variability)

6. Towards a complete theory of multisensory integration

Three decades ago, the question was posed whether there is a unified explanation for multisensory localization judgments under conflict (Warren, 1979). Behavioral theories of Bayes-optimal cue combination have brought us closer to this goal. Not only do they explain a wide range of existing data, they are also firmly rooted in a principled, probabilistic description of the purpose of multisensory perception, which is to increase precision if two cues have a common origin, but to keep cues with different origins segregated. They fit in a line of normative models of perception and decision-making that have been successful in recent years (Kording, 2007) and have a long history (Hatfield, 1990). However, whereas Bayesian approaches to behavioral data in multisensory perception have become commonplace, models for their underlying neural mechanisms have just started to appear. We believe that probabilistic population codes provide a framework that allows to build testable neural models of Bayes-optimal behavior without the need to make ad-hoc assumptions. Importantly, even when one or more of the assumptions are violated, the framework provides a systematic way to think about which corrections are needed.

In this review, we have focused on Bayesian cue integration and inference about the number of sources. From a computational view, there are two other aspects to multisensory perception: efficient read-out and coordinate transformations. Efficient read-out (decoding) is necessary to extract the largest possible amount of information from the multisensory population (Deneve et al., 1999), while coordinate transformations are necessary if the cues are initially not encoded in the same frame of reference, such as auditory (head-centered) and visual (eye-centered) cues (Deneve et al., 2001; Pouget and Sejnowski, 1997; Salinas and Abbott, 1995). Remarkably, if neuronal variability is Poisson-like, efficient read-out and coordinate transformations can be implemented through a basis function network with attractor dynamics (Deneve et al., 1999; 2001). A complete theory of multisensory integration should address at least all four aspects (see Table 1) and connect them in an overarching framework. It will only be possible to test such a theory through simultaneous measurements of behavior and of neural population activity.

Apart from a unifying theory of multisensory perception, there are still many interesting open questions in traditional cue integration. One of these is whether Bayesian models for cue integration can be applied to more complex stimuli, such as emotional state (De Gelder and Vroomen, 2000) or spoken syllables, either without (Ross et al., 2007; Sumby and Pollack, 1954) or with conflict (Massaro, 1987; 1998). Another outstanding issue concerns the relation between the statistics of the environment on the one hand and the likelihood function and priors in Bayesian models on the other hand. An intriguing new study shows that an arbitrary common-cause association between vision and touch can be learned with a moderate amount of training (Ernst, 2007). Moreover, people can learn to optimally take into the account the reliabilities of arbitrary low-level features of images (Michel and Jacobs, 2008). Finally, a topic of great interest is the time course of multisensory integration, which requires modeling of the full decision-making process and would allow to make contact with the literature on saccadic reaction times (Bell et al., 2005; Colonius and Arndt, 2001; Colonius and Diederich, 2004; Corneil et al., 2002; Diederich and Colonius, 2004; Diederich and Colonius, 2007; Rowland et al., 2007b).

REFERENCES

Alais, D., Burr, D., 2004. The ventriloquist effect results from near-optimal bimodal integration. Curr Biol 14, 257–262.

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.

Averbeck, B.B., Latham, P.E., Pouget, A., 2006. Neural correlations, population coding, and computation. Nat. Rev. Neurosci 7, 358–366.

Banks, M.S., 2004. What you see and hear is what you get. Curr. Biology 14, R236–238.

Battaglia, P.W., Jacobs, R.A., Aslin, R.N., 2003. Bayesian integration of visual and auditory signals for spatial localization. J. Opt. Soc. Am. A. Opt. Image Sci. Vis. 20, 1391–1397.

Beauchamp, M.S., 2005. Statistical criteria in fMRI studies of multisensory integration. Neuroinformatics 3, 93–113.

Beierholm, U., Kording, K., Shams, L., Ma, W.J., 2007. Comparing Bayesian models for multisensory cue combination without mandatory integration. In: Advances in Neural Information Processing Systems. Vol., ed.^eds.

Bell, A.H., Meredith, M.A., Van Opstal, A.J., Munoz, D.P., 2005. Crossmodal integration in the primate superior colliculus underlying the preparation and initiation of saccadic eye movements. J. Neurophysiol 93, 3659–3673.

Bertelson, P., Aschersleben, G., 1998. Automatic visual bias of perceived auditory location. Psychon. B. Rev 5, 482–489.

Bertelson, P., 1999. Ventriloquism: a case of crossmodal perceptual grouping. In: Ashersleben, G., Bachmann, T., Müsseler, J. (Eds.), Cognitive contributions to the perception of spatial and temporal events. Vol. Elsevier, Amsterdam.

Bertelson, P., Aschersleben, G., 2003. Temporal ventriloquism: crossmodal interaction on the time dimension 1. Evidence from auditory–visual temporal order judgment. Int J Psychophysiology 50, 147–155.

Bregman, A.S., 1990. Auditory Scene Analysis: The Perceptual Organization of Sound, Vol. MIT Press, Cambridge, MA.

Bresciani, J.P., Ernst, M.O., Drewing, K., Bouyer, G., Maury, V., Kheddar, A., 2005. Feeling what you hear: auditory signals can modulate tactile tap perception. Exp Brain Res. 162, 172–180.

Bresciani, J.P., Dammeier, F., Ernst, M.O., 2006. Vision and touch are automatically integrated for the perception of sequences of events. J Vis 6, 554–564.

Choe, C.S., Welch, R.B., Gilford, R.M., Juola, J.F., 1975. The "ventriloquist effect": visual dominance or response bias? Perception and Psychophysics 18, 55–60.

Clark, J.J., Yuille, A.L., 1990. Data Fusion for Sensory Information Processing Systems, Vol. Kluwer Academic, Boston.

Colonius, H., Arndt, P., 2001. A two-stage model for visual–auditory interaction in saccadic latencies. Percept Psychophys 63, 126–147.

Colonius, H., Diederich, A., 2004. Multisensory interaction in saccadic reaction time: a time-window-of-integration model. J Cogn Neurosci 16, 1000–1009.

Compte, A., Constantinidis, C., Tegner, J., Raghavachari, S., Chafee, M.V., Goldman-Rakic, P.S., Wang, X.J., 2003. Temporally irregular mnemonic persistent activity in prefrontal neurons

of monkeys during a delayed response task. J Neurophys 90, 3441–3454.
Corneil, B.D., Van Wanrooij, M., Munoz, D.P., Van Opstal, A.J., 2002. Auditory–visual interactions subserving goal-directed saccades in a complex scene. J Neurophysiol 88, 438–454.
De Gelder, B., Vroomen, J., 2000. The perception of emotion by ear and by eye. Cogn. Emot 14, 289–311.
De Gelder, B., Bertelson, P., 2003. Multisensory integration, perception and ecological validity. Trends in Cogn Sci 7, 460–467.
Dean, A.F., 1981. The variability of discharge of simple cells in the cat striate cortex. Exp Brain Res 44, 437–440.
Deneve, S., Latham, P., Pouget, A., 1999. Reading population codes: a neural implementation of ideal observers. Nature Neuroscience 2, 740–745.
Deneve, S., Latham, P., Pouget, A., 2001. Efficient computation and cue integration with noisy population codes. Nature Neuroscience 4, 826–831.
Diederich, A., Colonius, H., 2004. Bimodal and trimodal multisensory enhancement: effects of stimulus onset and intensity on reaction time. Percept Psychophys 66, 1388–1404.
Diederich, A., Colonius, H., 2007. Modeling spatial effects in visual-tactile saccadic reaction time. Percept Psychophys 69, 56–67.
Elder, J.H., Goldberg, R.M., 2002. Ecological statistics of Gestalt laws for the perceptual organization of contours. J. Vision 2, 324–353.
Ernst, M.O., Banks, M.S., 2002. Humans integrate visual and haptic information in a statistically optimal fashion. Nature 415, 429–433.
Ernst, M.O., 2007. Learning to integrate arbitrary signals from vision and touch. Journal of Vision 7 (5), 7.1–14.
Feldman, J., 2001. Bayesian contour integration. Percept. Psychophys 63, 1171–1182.
Feng, A.S., Ratnam, R., 2000. Neural basis of hearing in real-world situations. Annu. Rev. Psychol 51, 699–725.
Foldiak, P., 1993. The 'ideal homunculus': statistical inference from neural population responses. In: Eeckman, F., Bower, J. (Eds.), Computation and Neural Systems. Vol. Kluwer Academic Publishers, Norwell, MA, pp. 55–60.
Geisler, W.S., Perry, J.S., Super, B.J., Gallogly, D.P., 2001. Edge co-occurrence in natural images predicts contour grouping performance. Vision Res 41, 711–724.
Green, A.M., Angelaki, D.E., 2007. Coordinate transformations and sensory integration in the detection of spatial orientation and self-motion: from models to experiments. Progr. Brain Res 165, 155–180.
Gur, M., Beylin, A., Snodderly, D.M., 1997. Response variability of neurons in primary visual cortex (V1) of alert monkeys. J. Neurosci 17, 2914–2920.
Hatfield, G.C., 1990. The Natural and the Normative Theories of Spatial Perception from Kant to Helmholtz, Vol. MIT Press, Cambridge, Mass.
Holt, G.R., Softky, W.R., Koch, C., Douglas, R.J., 1996. Comparison of discharge variability in vitro and in vivo in cat visual cortex neurons. J. Neurophys. 75.
Howard, I.P., Templeton, W.B., 1966. Human Spatial Orientation, Vol. Wiley, New York.
Jack, C.E., Thurlow, W.R., 1973. Effects of degree of visual association and angle of displacement on the "ventriloquism" effect. Percept. Mot. Skills 37, 967–979.
Jacobs, R.A., 1999. Optimal integration of texture and motion cues to depth. Vision Res 39, 3621–3629.
Knill, D.C., 2003. Mixture models and the probabilistic structure of depth cues. Vision Res 43, 831–854.
Knill, D.C., Saunders, J.A., 2003. Do humans optimally integrate stereo and texture information for judgments of surface slant? Vision Research. 43, 2539–2558.
Kording, K., Tenenbaum, J.B., 2006. Causal inference in sensorimotor integration. In: Advances in Neural Information Processing Systems Vol., ed.^eds.
Kording, K., 2007. Decision theory: what "should" the nervous system do? Science 318, 606–610.
Kording, K.P., Beierholm, U., Ma, W.J., Quartz, S., Tenenbaum, J.B., Shams, L., 2007. Causal inference in multisensory perception. PLoS ONE. 2.
Latham, P.E., Deneve, S., Pouget, A., 2003. Optimal computation with attractor networks. Journal of Physiology (Paris) 97, 683–694.
Laurienti, P.J., Perrault, T.J., Stanford, T.R., Wallace, M.T., Stein, B.E., 2005. On the use of superadditivity as a metric for characterizing multisensory integration in functional neuroimaging studies. Exp Brain Res 166, 289–297.
Ma, W.J., Beck, J.M., Latham, P.E., Pouget, A., 2006. Bayesian inference with probabilistic population codes. Nat Neurosci 9, 1432–1438.
Mamassian, P., 2006. Bayesian inference of form and shape. Progr. Brain Res. 154, 265–270.
Massaro, D.W., 1987. Speech Perception by Ear and Eye: A Paradigm for Psychological Inquiry, Vol. Lawrence Erlbaum Associates, Hillsdale, NJ.
Massaro, D.W., 1998. Perceiving Talking Faces: From Speech Perception to a Behavioral Principle, Vol. MIT Press, Cambridge, MA.
Meredith, M.A., Stein, B.E., 1986. Spatial factors determine the activity of multisensory neurons in cat superior colliculus. Cogn Brain Res 369, 350–354.
Michel, M.M., Jacobs, R.A., 2008. Learning optimal integration of arbitrary features in a perceptual discrimination task. Journal of Vision 8 (2), 3.1–16.
Nirenberg, S., Latham, P.E., 2003. Decoding neuronal spike trains: how important are correlations? Proc Natl Acad Sci U S A. 100, 7348–7353.
Patton, P., Anastasio, T.J., 2003. Modeling cross-modal enhancement and modality-specific suppression in multisensory neurons. Neural Computation 15, 783–810.
Perrault, T.J., Vaughan, J.W., Stein, B.E., Wallace, M.T., 2005. Superior colliculus neurons use distinct operational modes in the integration of multisensory stimuli. J Neurophysiol 93, 2575–2586.
Populin, L.C., Yin, T.C., 2002. Bimodal interactions in the superior colliculus of the behaving cat. J Neurosci 22, 2826–2834.
Pouget, A., Sejnowski, T., 1997. Spatial transformations in the parietal cortex using basis functions. Journal of Cognitive Neuroscience 9, 222–237.
Radeau, M., Bertelson, P., 1987. Auditory–visual interaction and the timing of inputs. Psychol Res 49, 17–22.
Roach, N.W., Heron, J., McGraw, P.V., 2006. Resolving multisensory conflict: a strategy for balancing the costs and benefits of audio-visual integration. Proc Biol Sci 273, 2159–2168.
Ross, L.A., Saint-Amour, D., Leavitt, V.N., Javitt, D.C., Foxe, J.J., 2007. Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments. Cereb. Cortex 17, 1147–1153.
Rowland, B., Stanford, T., Stein, B.E., 2007a. A Bayesian model unifies multisensory spatial localization with the physiological properties of the superior colliculus. Exp. Brain Res 180, 153–161.
Rowland, B.A., Quessy, S., Stanford, T.R., Stein, B.E., 2007b. Multisensory integration shortens physiological response latencies. J. Neurosci 27, 5879–5884.
Sahani, M., Dayan, P., 2003. Doubly distributional population codes: simultaneous representation of uncertainty and multiplicity. Neural. Comput 15, 2255–2279.
Salinas, E., Abbott, L., 1995. Transfer of coded information from sensory to motor networks. Journal of Neuroscience 15, 6461–6474.
Sanger, T., 1996. Probability density estimation for the interpretation of neural population codes. Journal of Neurophysiology 76, 2790–2793.

Sato, Y., Toyoizumi, T., Aihara, K., 2007. Bayesian inference explains perception of unity and ventriloquism aftereffect: identification of common sources of audiovisual stimuli. Neural Computation 19, 3335–3355.
Series, P., Latham, P., Pouget, A., 2004. Tuning curve sharpening for orientation selectivity: coding efficiency and the impact of correlations. Nature Neuroscience. 10, 1129–1135.
Shams, L., Ma, W.J., Beierholm, U., 2005. Sound-induced flash illusion as an optimal percept. Neuroreport 16, 1923–1927.
Slutsky, D.A., Recanzone, G.H., 2001. Temporal and spatial dependency of the ventriloquism effect. Neuroreport 12, 7–10.
Stanford, T.R., Quessy, S., Stein, B.E., 2005. Evaluating the operations underlying multisensory integration in the cat superior colliculus. J. Neurosci 25, 6499–6508.
Stanford, T.R., Stein, B.E., 2007. Superadditivity in multisensory integration: putting the computation in context. Neuroreport 18, 787–791.
Stein, B.E., Huneycutt, W.S., Meredith, M.A., 1988. Neurons and behavior: the same rules of multisensory integration. Brain Res 448, 355–358.
Stein, B.E., Meredith, M.A., 1993. The Merging of the Senses. MIT Press, Cambridge, MA.
Sumby, W.H., Pollack, I., 1954. Visual contribution to speech intelligibility in noise. J.Acoust. Soc. Am. 26, 212–215.
Thomas, G.J., 1941. Experimental study of the influence of vision on sound localization. J Exp Psychol 28, 163–177.
Thurlow, W.R., Jack, C.E., 1973. Certain determinants of the "ventriloquism effect". Percept Mot Skills 36, 1171–1184.
Tolhurst, D., Movshon, J., Dean, A., 1982. The statistical reliability of signals in single neurons in cat and monkey visual cortex. Vision Research 23, 775–785.
Treue, S., Hol, K., Rauber, H., 2000. Seeing multiple directions of motion-physiology and psychophysics. Nature Neuroscience 3, 270–276.
van Beers, R.J., Sittig, A.C., Denier van der Gon, J.J., 1996. How humans combine simultaneous proprioceptive and visual position information. Exp. Brain Res 111, 253–261.
Vroomen, J., Keetels, M., 2006. The spatial constraint in intersensory pairing: no role in temporal ventriloquism. J. Exp. Psychol. Hum. Percept. Perform 32, 1063–1071.
Wallace, M.T., Wilkinson, L.K., Stein, B.E., 1996. Representation and integration of multiple sensory inputs in primate superior colliculus. J Neurophysiol 76, 1246–1266.
Wallace, M.T., Roberson, G.E., Hairston, W.D., Stein, B.E., Vaughan, J.W., Schirillo, J.A., 2004. Unifying multisensory signals across time and space. Exp Brain Res 158, 252–258.
Warren, D.H., 1979. Spatial localization under conflict conditions: Is there a single explanation? Perception. 8, 323–337.
Welch, R.B., 1972. The effect of experienced limb identity upon adaptation to simulated displacement of the visual field. Perception and Psychophysics 12, 453–456.
Welch, R.B., Warren, D.H., 1980. Immediate perceptual response to intersensory discrepancy. Psychol Bull 88, 638–667.
Wolpert, D., Ghahramani, Z., Jordan, M., 1995. An internal model for sensorimotor integration. Science 269, 1880–1882.
Yuille, A.L., Bulthoff, H.H., 1996. Bayesian decision theory and psychophysics. In: Knill, D.C., Richards, W. (Eds.), Perception as Bayesian Inference. Vol. University Press, New York: Cambridge.