Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Physiology - Paris 98 (2004) 249–258
Journal of Physiology Paris
www.elsevier.com/locate/jphysparis

Bayesian multisensory integration and cross-modal spatial links

Sophie Deneve [a,*], Alexandre Pouget [b]

[a] Gatsby Computational Neuroscience Unit, Alexandra House, 17 Queen Square, London WC1N 3AR, UK
[b] Department of Brain and Cognitive Sciences, Meliora Hall, University of Rochester, Rochester, NY 14627, USA

Abstract

Our perception of the word is the result of combining information between several senses, such as vision, audition and proprioception. These sensory modalities use widely different frames of reference to represent the properties and locations of object. Moreover, multisensory cues come with different degrees of reliability, and the reliability of a given cue can change in different contexts. The Bayesian framework—which we describe in this review—provides an optimal solution to deal with this issue of combining cues that are not equally reliable. However, this approach does not address the issue of frames of references. We show that this problem can be solved by creating cross-modal spatial links in basis function networks. Finally, we show how the basis function approach can be combined with the Bayesian framework to yield networks that can perform optimal multisensory combination. On the basis of this theory, we argue that multisensory integration is a dialogue between sensory modalities rather that the convergence of all sensory information onto a supra-modal area.

Keywords: Multisensory integration; Basis functions; Recurrent networks; Bayesian; Frames of reference; Gain modulation; Partially shifting receptive fields

1. Introduction

Multisensory integration refers to the capacity of combining information coming from different sensory modalities to get a more accurate representation of the environment and body. For example, vision and touch can be combined to estimate the shape of objects, and viewing somebody's lips moving can improve speech comprehension. This integration process is difficult for two main reasons. First, the reliability of sensory modalities varies widely according to the context. For example, in daylight, visual cues are more reliable than auditory cues to localize objects, while the contrary is true at night. Thus, the brain should rely more on auditory cues at night and more on visual cues during the day to estimate object positions.

Another reason why multisensory integration is a complex issue is that each sensory modality uses a different format to encode the same properties of the environment or body. Thus multisensory integration cannot be a simple averaging between converging sensory inputs. More elaborate computations are required to interpret neural responses corresponding to the same object in different sensory areas. To use an analogy in the linguistic domain, each sensory modality uses its own language, and information cannot be shared between modalities without translation mechanisms for the different languages. For example, sensory modality encodes the position of objects in different frames of reference. Visual stimuli are represented by neurons with receptive fields on the retina, auditory stimuli by neurons with receptive fields around the head, and tactile stimuli by neurons with receptive fields anchored on the skin. Thus, a change in eye position or body posture will result in a change in the correspondence between visual, auditory and tactile neural responses encoding the same object. To combine these different sensory responses, the brain must take into account the posture and the movements of the body in space.

We first review the Bayesian framework for multisensory integration, which provides a set of rule to optimally combine sensory inputs with varying reliabilities. We then describe several psychophysical studies

[*] Corresponding author.
E-mail addresses: sdeneve@gatsby.ucl.ac.uk (S. Deneve), alex@bcs.rochester.edu (A. Pouget).

0928-4257/$ - see front matter © 2004 Published by Elsevier Ltd.
doi:10.1016/j.jphysparis.2004.03.011

supporting the notion that multisensory integration in the nervous system is indeed akin to a Bayesian inference process. We then review evidence from psychophysics and neuropsychology that sensory inputs from different modalities, but originating at the same location in space, can influence one another regardless of body posture, suggesting that there is a link, or translation mechanism, between the spatial representations of different sensory systems. Finally, we turn to neurophysiological and modeling data regarding the neural mechanisms of spatial transformations and Bayesian inferences.

2. Bayesian framework for multisensory integration

The Bayesian framework allows the optimal combinations of multiple sources of information about a quantity x [22]. We consider a specific example in which x refers to the position of an object which can be seen and heard at the same time. Given noisy neural responses in the visual cortex, $\mathbf{r}_{\text{vis}}$, the position of the object is most probably near the receptive fields of the most active cells, but this position cannot be determined with infinite precision due to the presence of neural noise (we use bold letter to refer to vector; thus $\mathbf{r}_{\text{vis}}$ is meant to be a vector corresponding to the firing rate of a large population of visual neurons). Given the uncertainty associated with x, a good strategy is to compute the posterior probability that the object is at position x given the visual neural responses, $P(x|\mathbf{r}_{\text{vis}})$. Using Baye's rule, $P(x|\mathbf{r}_{\text{vis}})$ can be obtained by combining the distribution of neural noise $P(x|\mathbf{r}_{\text{vis}})$ with prior knowledge on the distribution of object position $P(x)$ and the prior probability of neural responses $P(\mathbf{r}_{\text{vis}})$:

$$P(x|\mathbf{r}_{\text{vis}}) = \frac{P(\mathbf{r}_{\text{vis}}|x)P(x)}{P(\mathbf{r}_{\text{vis}})} \tag{1}$$

This distribution is called a posterior probability because it refers to the probability of object position *after* taking into account the sensory input, $\mathbf{r}$ (as opposed to the prior $P(x)$ which is independent of $\mathbf{r}$). $P(x|\mathbf{r}_{\text{vis}})$ is called the noise distribution because it corresponds to variability in neural responses for a fixed stimulus, i.e., to variability which is not accounted by the stimulus.

Note several important points about Eq. (1). First, we can ignore the denominator, $P(\mathbf{r}_{\text{vis}})$, because it is independent of x, and x is the only variable we care about. Second, $P(\mathbf{r}_{\text{vis}}|x)$ can be measured experimentally by repetitively presenting an object at the same position x and measuring the variability in $\mathbf{r}_{\text{vis}}$ (which is why we call this term the "noise" distribution). Finally, if we happen to know that the object is more likely to appear in some visual locations than others, we can represent this knowledge in the prior distribution $P(x)$. In this section, we will assume that all positions are equally likely, that is, $P(x) = c$, where c is a constant. This implies that $P(x)$ does not depend on x, in which case we can also ignore it. Therefore, Eq. (1) reduces to:

$$P(x|\mathbf{r}_{\text{vis}}) \propto P(\mathbf{r}_{\text{vis}}|x) \tag{2}$$

Once the posterior distribution is computed, an estimate of the position of the object can be obtained by recovering the value of x that maximizes that distribution:

$$\hat{x}_{\text{vis}} = \arg\max_x P(x|\mathbf{r}_{\text{vis}})$$

This is known as the maximum a posteriori estimate, or MAP estimate for short.

When we hear the object, a similar posterior distribution, $P(x|\mathbf{r}_{\text{aud}})$, and its corresponding estimate, $\hat{x}_{\text{aud}}$, can be computed based on the noisy responses of auditory neuron, $\mathbf{r}_{\text{aud}}$.

What should we do when the object is heard and seen at the same time? Using the same approach, we need to compute the estimate, $\hat{x}_{\text{bim}}$ ("bim" stands for bimodal), maximizing the posterior distribution, $P(x|\mathbf{r}_{\text{vis}}, \mathbf{r}_{\text{aud}})$:

$$\hat{x}_{\text{bim}} = \arg\max_x P(x|\mathbf{r}_{\text{vis}}, \mathbf{r}_{\text{aud}})$$

To compute the posterior distribution, we use Bayes law, which under the assumption of a flat prior distribution, reduces to:

$$P(x|\mathbf{r}_{\text{vis}}, \mathbf{r}_{\text{aud}}) \propto P(\mathbf{r}_{\text{vis}}, \mathbf{r}_{\text{aud}}|x) \tag{3}$$

$$P(x|\mathbf{r}_{\text{vis}}, \mathbf{r}_{\text{aud}}) \propto P(\mathbf{r}_{\text{vis}}|x)P(\mathbf{r}_{\text{aud}}|x) \tag{4}$$

$$P(x|\mathbf{r}_{\text{vis}}, \mathbf{r}_{\text{aud}}) \propto P(x|\mathbf{r}_{\text{vis}})P(x|\mathbf{r}_{\text{aud}}) \tag{5}$$

To go from Eqs. (3) and (4), we assumed that the noise corrupting the visual neurons is independent from the one corrupting the auditory neurons (which seems reasonable given how far apart those neurons are in the cortex). The step from Eqs. (4) and (5) is a consequence of Eq. (2). From Eq. (5), we see that the bimodal posterior distribution can be obtained by simply taking the product of the unimodal distributions. An example of this operation is illustrated in Fig. 1.

When $P(x|\mathbf{r}_{\text{vis}})$ and $P(x|\mathbf{r}_{\text{aud}})$ are Gaussian probability distributions, as is the case in Fig. 1, the bimodal estimate $\hat{x}_{\text{bim}}$ can be obtained by taking a linear combination of the unimodal estimates, $\hat{x}_{\text{vis}}$ and $\hat{x}_{\text{aud}}$, weighted by their respective reliabilities [6,21,41]:

$$\hat{x}_{\text{bim}} = \frac{1/\sigma^2_{\text{vis}}}{1/\sigma^2_{\text{vis}} + 1/\sigma^2_{\text{aud}}}\hat{x}_{\text{vis}} + \frac{1/\sigma^2_{\text{aud}}}{1/\sigma^2_{\text{vis}} + 1/\sigma^2_{\text{aud}}}\hat{x}_{\text{aud}}, \tag{6}$$

where $1/\sigma^2_{\text{vis}}$ and $1/\sigma^2_{\text{aud}}$, the reliability of the visual and auditory estimates respectively, are the inverse of the variances of the visual and auditory posterior probabilities. In particular, if the visual input is more reliable than the auditory input (σ^2_{vis} is smaller than σ^2_{aud}) then the bimodal estimate of position should be closer to the visual estimate and vice versa if audition is more reliable than vision.

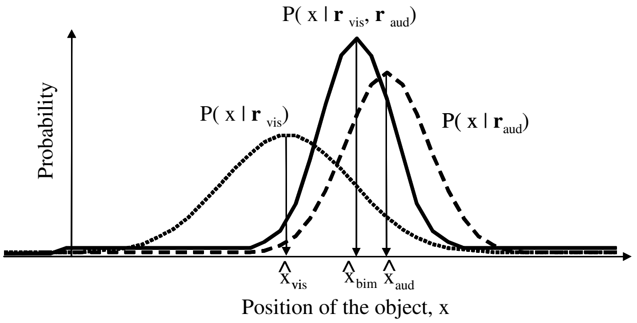

Fig. 1. Posterior probability of an object position given the visual input (dotted line), the auditory input (dashed line), and the visual and auditory inputs combined (solid line). $\hat{x}_{\text{vis}}$, $\hat{x}_{\text{aud}}$ and $\hat{x}_{\text{bim}}$ correspond to the MAP (maximum a posteriori) estimate of the position of the object based on the visual input alone, the auditory input alone and the combination of both inputs.

Can the brain employ such adaptive procedure to combine optimally neural responses in different sensory modalities? Does it take into account the relative reliability of different sensory cues before combining them? Recent psychophysical data suggests that this might indeed be the case.

3. Evidence for Bayesian multisensory integration

One method for studying multisensory integration is to compare the distribution of estimates made by human subjects from unimodal and bimodal sensory inputs. The Bayesian hypothesis predicts that the distribution of bimodal estimates should be approximately a product between the unimodal estimate distributions (Eq. (5)).

This approach has been applied successfully to the estimated position of the hand from visual and proprioceptive inputs [33–35]. In these experiments, subjects were required to localize a proprioceptive target (their middle finger of their right hand without visual feedback), a visual target, or a visual and proprioceptive target (their middle finger of their right hand with visual feedback). They indicated the estimated target position by pointing with their left hand (see Fig. 2a). The distributions of position estimate for unimodal visual or proprioceptive targets are anisotropic in space, but their axes of maximum variance are oriented differently. As schematized on Fig. 2b, vision is most reliable in azimuth, but less reliable in depth, i.e. on a radial direction compare to the observer. On the other hand, proprioception is more reliable in depth than in azimuth. Moreover, visual and proprioceptive estimates have subject-specific biases, that is, they are systematically deviated from the true target position (represented by a red circle on Fig. 2b). As a consequence, the bimodal estimate predicted by the Bayesian framework—a product between the unimodal distributions, Eq. (5)—should be less variable than the unimodal estimates. Moreover it should, on average, be deviated from to the straight line between visual and proprioceptive mean estimates. This is because the bimodal estimate lies near the intersection between the axes of maximum variance of the visual and proprioceptive estimates, as illustrated

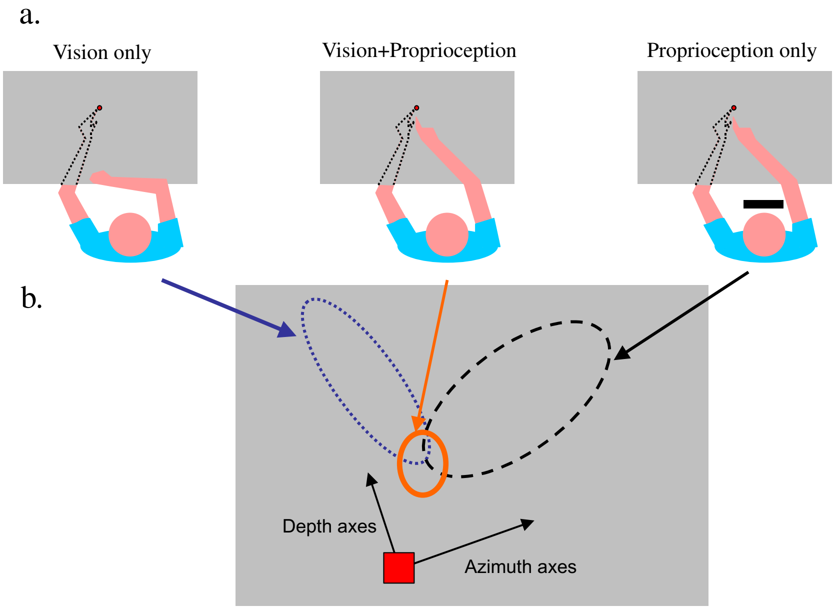

Fig. 2. Bayesian integration of visual and proprioceptive cues. (a) Experimental setting (see text) for van Beers et al. experiments. Subjects sitting at a table were required to point with their left hand under the table to the middle finger of their right hand and/or a visual target. (b) Schematic distribution of visual (dotted line), proprioceptive (dashed line) and bimodal (solid line) estimates of target position. The true position of the object is represented by the red circle. Adapted from van Beers et al. [35]. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

in Fig. 2b. These two predictions were confirmed by experimental data [33–35]. Thus, this is evidence that the brain performs an optimal integration process and takes into account the specific error distribution associated with each modality before combining them.

Recently, van Beers et al. provided further evidence for this claim [36]. They exposed their subjects to an adaptation period during which the visual feedback was systematically displaced compare to the true finger position. Visual displacements could occur along two different axes: the azimuth axes or the depth axes. They measured the after-effect of adaptation by asking their subjects to localize visual or proprioceptive targets before and after the adaptation period, using the experimental setting described above. The relative strength of adaptation in each sensory modality is an indirect measurement of the weights given to each in an integrated percept of hand position [29,40]. According to the Bayesian framework, these weights should be proportional to the reliability of each sensory modality along the axes of displacement used during adaptation. Thus vision, which is the most reliable modality for azimuth, should be given a stronger weight and present less after-effect when adaptation occurs along the azimuth axes. The reverse should be true along the depth axes where proprioception is the most reliable modality. This is indeed what was observed [36].

However, these axes-dependent weights of the two sensory cues could be due to pre-wired characteristics of visual and postural neural networks, without regard to the actual context-dependent reliabilities. It remains to be shown that the brain can adapt on-line (i.e. from trial to trial) to change in the reliability of each sensory cue.

Two other groups [3,15] have also collected indirect evidence for Bayesian inferences in multisensory integration. Their primary task was a visual cue combination task, that is, the combination of two visual cues, like texture and disparity, to infer the 3D properties of objects. However, they used another sensory modality, haptic feedback, to probe the reliability of each visual cue. As we briefly described in the introduction, a Bayesian estimate based on two sensory inputs with normal noise distributions should be a linear combination of the unimodal estimates (Eq. (6)), weighted by the reliability of each sensory cue. If haptic feedback is systematically consistent with one visual cue and inconsistent with the other, it should contribute to increase the perceived reliability of the consistent cue and decrease the reliability of the inconsistent cue. For example, if texture is consistent with the haptic percept of surface orientation but disparity is not, texture should be perceived as more reliable and given a stronger weight in estimating surface orientation. These experiments showed indeed that after an adaptation period during which haptic feedback is provided, subject increased the weight given to the cue consistent with haptic feedback and decreased the weight of the inconsistent cue (while being reportedly unaware that one of the cue is inconsistent with the haptic percept). Thus, multisensory integration can be used to calibrate visual inference processes in a context-dependent fashion. However, these experiments rely on an intensive training of the subjects, and are only an indirect measurement of multisensory integration.

A more direct evidence for an adaptive, context-dependent process in multisensory integration comes from a recent experiment by Ernst and Banks [14]. The task was to discriminate the thickness of bars presented visually (random dot stereogram) or through haptic feedback. They measured the discrimination thresholds of their subjects while varying the reliability of the visual cue from trial to trial (the reliability of the haptic cue was maintained constant). Bayesian integration of the two sensory cues should result in a linear combination of visual and haptic thickness with weights proportional to their respective reliabilities. Ernst and Banks used bimodal visual/haptic bars with discrepancies between the visual and haptic thickness to measure the weights given to the visual modality in different visual noise conditions (Fig. 3a). They also measured the discrimination threshold for bimodal bars (Fig. 3b). The weights and bimodal discrimination thresholds turned out to be very close to the ones predicted by the Bayesian framework (Fig. 3a and b, dashed lines). Thus, humans appear to be able to combine visual and haptic cues optimally, adapting their strategy to the reliability of each cue.

Unfortunately, the Bayesian model says nothing about the neural mechanism by which such integration could be performed. It is still unclear how reliabilities and probability distributions could be represented in

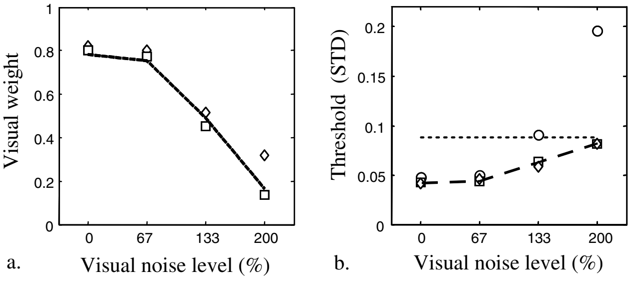

Fig. 3. Bayesian integration of visual and haptic cues. (a) Weight given to the visual modality in estimating the width of a bar from bimodal visual and haptic sensory inputs. The visual weight is plotted for four different levels of noise added to the visual input. Diamond: experimental measurements [14]. Dashed line: predictions from the Bayesian model. Squares: Visual weights from the iterative basis function network [8]. (b) Discrimination thresholds for the width of visual bars (circles), haptic bars (dotted line) and bimodal visual and haptic bars (diamond) [14]. Dashed line: Bimodal discrimination thresholds predicted from the Bayesian model. Squares: Bimodal discrimination thresholds from the iterative basis function network.

neural networks. In particular it does not solve the spatial sensory correspondence issue: Sensory representations are in different frames of reference and to be integrated they need to be compared with the posture. To do so, the brain must be able to link the body-centered positions in all sensory modalities corresponding to the same position in space. We will review evidences that such a spatial link exist in humans and non-human primates between vision, audition and the tactile modality before showing how it can be combined with the Bayesian approach.

4. Spatial links between sensory modalities

The existence of a spatial links between sensory modalities is supported by the existence of spatially selective cross-modal attentional effects. If the representations of space issued from different sensory modalities are ultimately combined, one might expect that exogenous attention attracted by a visual stimulus could facilitate tactile detection or discrimination at this location, and vice-versa. These cross-modal attentional effects have been observed for all combinations of sensory modalities [10,39]. A case of particular interest is when the alignment between different sensory frames of reference is systematically modified. For example, changing the eye position without moving the head change the alignment between visual (eye-centered) and auditory (head-centered) frames of reference, and crossing the hands change the alignment between visual and somatosensory (skin-centered) frames of reference. In all these conditions, the cross-modal attentional effects were remapped appropriately between sensory modalities. Thus a visual stimulus near one hand attracts tactile attention on this hand, even when the arms are crossed, and a salient visual stimulus attracts auditory attention at that location regardless of eye position [9]. To implement such attentional spatial links, cortical networks have to take into account the posture and perform coordinate transform between eye-centered, skin-centered and head-centered frames of reference.

Cross-modal attentional effects can also be investigated in human patients with cortical lesions, and in particular neglect patients. These patients have unilateral brain lesion, usually in the right parietal or frontal cortex, and have difficulties detecting or responding to stimuli in the contralesional space. In particular, they often exhibit extinction: when two stimuli are represented simultaneously, the leftward stimulus, which would be detected if presented in isolation, is extinguished by the rightward stimuli and only the rightward stimulus is reported by the patient. In numerous cases, this extinction is not restricted to the visual modality but extend to the tactile and auditory modality as well. Interestingly, extinction can also occur between visual and tactile events, that is, a visual stimulus near the right hand extinguish a tactile stimulus on the left hand and vice-versa [24,27]. This effect is spatially selective, and reduced, but not suppressed, when the visual and tactile stimuli are in non-homologue locations, that is, when the visual stimulus is not in the close vicinity of the right hand. Auditory-tactile extinction in the near-space around the head has also been reported [23]. These extinction experiments are strong evidence that cross-modal spatial attentional effects exist and that coordinate transforms are performed in cortical networks.

What could be the neural mechanism implementing this cross-modal spatial link? Two main hypothesis can be contrasted: Sensory remapping, which would involve the recoding of all sensory inputs in a common frame of reference on a multisensory brain area, and direct cross-modal influence, whereby sensory activity in one unimodal brain area directly influences sensory activities in another unimodal area. To use once again the linguistic metaphor, sensory remapping would consist in translating all modality-specific languages in a supra-modal language, while sensory modalities would be unable to communicate directly with one another. Direct cross-modal influences would consist in having a set of rules allowing each modality to translate the information contained in the other sensory systems in its own language and thus share information directly without a common language. In an attempt to identify brain areas involved in cross-modal spatial interactions, Macaluso et al. used brain imaging studies while subject where presented with lateralized visual, tactile and bimodal stimuli [25]. Not surprisingly, they found unimodal areas that responded only to tactile (post central gyrus) or visual (lateral and inferior occipital lobe) contralateral stimuli. They also found multisensory areas activated by both contralateral visual and tactile stimuli (anterior intra-parietal sulcus). This first result supports the sensory remapping hypothesis, according to which inputs from unimodal sensory areas would converge on multisensory areas to be presumably recoded in the same frame of reference.

However, in the same studies the authors also observed that the activity in unimodal visual areas in response to visual stimuli is enhanced by a simultaneous tactile stimulation at the same location in space. Moreover, this cross-modal activation remap appropriately with the posture: When fixation is straight ahead, a tactile stimulus enhances visual responses in the visual cortex contralateral to the tactile stimulation site. However, when fixations is at a peripheral location, so that a visual stimulus appear on the side of the retina opposite to the side of the tactile stimulation, tactile stimuli enhance activity in the visual cortex *ipsilateral* to the tactile stimulation site [26]. Similarly, spatially selective cross-modal influences between visual, tactile and auditory sensory responses have been found in ERP studies. These effects persist and remap appropriately when the hands are crossed, but the cross-modal enhancement in this condition is weaker [13,20]. These results show that cross-modal spatial links in attention can also be implemented through modulation of the unimodal sensory responses by inputs from the other sensory modalities, in support of the direct cross-modal influence hypothesis.

The existence of zones of sensory convergence (as represented by areas responding to stimuli in all sensory modalities) as well as cross-modal influences on unimodal brain areas suggests a role of both feed-forward connections from unimodal to multimodal areas and feedback connections from multimodal to unimodal areas [11]. Later in the next section, we will present a model of multisensory integration at the neural level which implements both processing streams and use them to perform optimal multisensory integration.

5. Neural implementation of multisensory Bayesian inference

We now turn to models that have attempted to tackle the neural processes involved in Bayesian multisensory integration. As these models find support in neurophysiological data, we will also report experimental results obtained by recording from multisensory cells in non-human subjects.

The Bayesian framework has been recently used by Anastasio et al. [1] to interpret response of deep layer superior colliculus (SC) cells to unimodal and bimodal stimuli. Some of these cells respond to visual, tactile and auditory stimuli and present a phenomenon called multisensory enhancement, whereby responses to bimodal stimuli are stronger than the response to the best unimodal stimulus [30,37,38]. This enhancement is considered as a signature of multisensory integration, and disappears when visual and auditory stimuli are not congruent in space or time. This parallels existing behavioral effects in cat [31] and in man [16], whereby auditory targets facilitate responses to visual target at the same position in space. Moreover, this enhancement is stronger for weaker stimuli, a phenomenon referred to as inverse effectiveness. Anastasio et al. [1] showed that multisensory enhancement and inverse effectiveness could be explained if SC cells represent the probability of a target being present in their receptive field given their visual and auditory input. According to this model, the multisensory enhancement observed in collicular cells and reflected in behavior would be due to an increase in the probability of a target being present when two sensory inputs are available, compare to a single one. However, this increase is important when the best unimodal sensory input is weak and, by consequent, another sensory input will significantly increase the probability of the target being present. If the best unimodal stimulus is strong, the probability of the target being present is already 1 and the response of the cell cannot increase much further. This, in turn, result in the observed inverse effectiveness.

SC cell have localized receptive fields and encode the position of a stimulus. Thus, a natural extension of Anastasio et al. approach is to propose that these cells encode the probability of a target being present *at position x* given their visual and/or auditory input, x being the position of the cell's receptive field. Under this assumption, we can make a parallel between the predicted activity of the population of SC cells and the posterior probabilities on object position useful for Bayesian inference. If we plot the visual responses of these model cells as a function of the position of their receptive, x, we would get $P(x|R_{\text{vis}})$, the dotted curve in Fig. 1a. Similarly, the auditory responses would correspond to $P(x|R_{\text{aud}})$, the dashed line on Fig. 1a. And for bimodal visual and auditory stimuli, the bimodal responses would correspond to the combined posterior probability $P(x|R_{\text{vis}}, R_{\text{aud}}) = P(x|R_{\text{vis}})P(x|R_{\text{aud}})$, the solid line in Fig. 1a. The estimated object position could be read directly from the neural population as the position of the peak activity. One advantage of this model is that it naturally implements Bayesian inference without requiring a specific mechanism devoted to it. The neural responses to bimodal stimuli would simply have to be the product between visual and auditor neural responses [14]. In a more general context, it has been proposed that cortical neurons representing sensory and motor variables with population codes might implement filtered version of the posterior probability distribution for these variables (Fig. 4a) [42].

However there are two main problems with applying these models to multisensory integration. First of all, increasing the information present in the sensory inputs (by increasing the strength or duration of the stimuli, for example) will result is a sharper distribution for $P(x|R_{\text{vis}}, R_{\text{aud}})$. At the extreme, when the sensory inputs are extremely reliable, only one object location will be highly probable (infinite precision) and the distribution of activity, corresponding to $P(x|R_{\text{vis}}, R_{\text{aud}})$, will be zero everywhere except at the correct location. As a consequence, SC cells encoding location close to the true object position will have activities that *decrease* when the sensory input strength or duration increases (see the cells circles in dotted line on Fig. 3a). Given that there is a direct relationship between the width of the distribution of activity on the neural population and the size of the receptive fields, this also means that receptive fields should decrease in size for more reliable sensory inputs. This is clearly not true in some situation, such as when contrast is being manipulated. The reliability of a visual stimulus is proportional to its contrast, yet the width of orientation or spatial frequency tuning and the size of

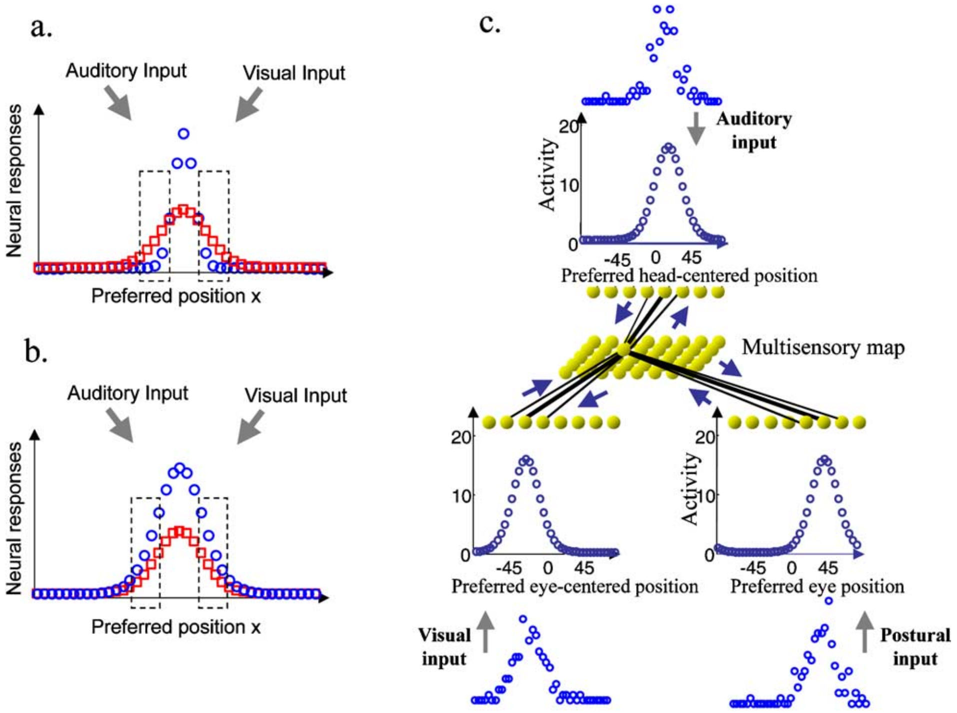

Fig. 4. Models of Bayesian multisensory integration. (a) Activity of a population of neurons encoding explicitly the posterior probability of the position of an object given visual and auditory inputs, $P(x|R_{\mathrm{vis}}, R_{\mathrm{aud}})$. The blue circles correspond to a narrow posterior distribution, i.e., a reliable sensory input, while the red squares show the activity pattern for a wide posterior, i.e., an unreliable sensory input. Cells within the dotted square decrease their activity when the reliability of the sensory input increases. (b) Activity of a population of cells representing the position of an object with fixed tuning curves (fixed receptive fields), but whose response gain vary with the reliability of the sensory inputs. In this case, the activity of all cells grows with the reliability of the sensory input. (d) Iterative basis function network performing multi-directional sensory predictions. Noisy sensory inputs are clamped onto the visual, auditory and postural input layers, after which the network converges to stable hills of activities. The positions of the stable hills are the network estimates for the position of the object in eye-centered and head-centered coordinates, as well as the position of the eyes (adapted from Deneve et al. [8]). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

the receptive fields of visual neurons do not change with contrast. Instead, the increase of reliability is reflected by a higher response gain (Fig. 3b).

Another problem with these probabilistic models is that they are incomplete as a theory of spatial multisensory integration. The positions represented by unimodal visual and auditory neural responses are not the same: Visual neurons code for eye-centered positions while auditory neurons code for head-centered positions. A third parameter, eye position, which links these two frames of reference, must be introduced in the integration process.

One possible solution to this last problem is to propose that sensory responses are remapped in a common frame of reference before converging on multisensory cells, so that all sensory input to multi-modal cells are in the same format. There are evidences from electrophysiological recordings in monkeys that sensory input tends to be recoded in the same frames of reference on multisensory areas. Thus, auditory targets for saccades appear to be remapped in an eye-centered frame of reference in the SC [19], and auditory targets for reaching are encoded in an eye-centered frame of reference in the parietal reach region [5,7]. Similarly, visual, auditory and tactile stimuli are remapped in a skin-centered frame of reference in the premotor cortex [17,18]. However, this approach does not solve the problem of *how* auditory inputs are remapped in eye-centered frames of reference or visual inputs in skin-centered frame of reference. Moreover, the coordinate transforms do not appear to be complete in many cases. For example in Jay and Sparks data, 1987, often cited as an evidence for the remapping of auditory cue in an eye-centered frame of reference in SC, the average auditory shift with the eye is only 50% of the total gaze shift, and cells are gain modulated by eye position. Further evidence for partially shifting receptive fields are auditory target for memorized saccades in lateral intra-parietal area LIP [32], visual receptive fields in VIP [12], and visual targets for reaching in area 5 [7].

Recently we proposed a model of Bayesian multisensory integration in cortical networks that accounts for the spatial links between sensory modalities [8]. These networks essentially perform multi-directional sensory predictions, using basis function maps (multi-modal areas) as a computational intermediate. According to this view, both cross-modal spatial links and Bayesian integration are two sides of the same coin. In both cases, the essential computation is the capacity to predict the outcome of an event in one sensory modality from the corresponding input in other sensory modalities and the posture. The role of multisensory areas is to compute the necessary coordinate transforms. In the linguistic metaphor, multisensory brain areas are the dictionaries allowing each sensory modality to translate information from all other modalities in its own language.

The architecture of the network we used to illustrate the theory is represented on Fig. 3c. It consist in 3 input layers, a visual layer coding for the eye-centered position of an object, an auditory layer coding for the head-centered position of an object and a postural layer coding for eye position. The population codes in the three input layers correspond to the responses of a population of visual cells with eye-centered receptive fields, a population of auditory cells with head-centered receptive fields and a population of postural cells with eye position gain fields. Note that, for symmetry purposes, we used Gaussian rather than sigmoid gain fields in the postural layer. Gain fields are usually described as being monotonic rather that bell-shaped [2]. However, our network can easily be extended to monotonic gain fields without modifying any of its interesting properties. The input units do not represent posterior probability but, more realistically, code for object position with fixed tuning curves, corresponding to the shape of their receptive fields. These unimodal input layers are interconnected with a multisensory intermediate layers with basis function units. At each iteration, the eye-centered (visual), head-centered (auditory) and eye position (postural) inputs are combined on the multisensory layer. These multisensory activities are then fed back into the input layers, in a way that compute the eye-centered position from the head-centered position and eye position, and vice versa the head-centered position from the eye-centered and eye position. This process is iterated until the network stabilizes, that is, until an agreement is reached between the visual and auditory position encoded in the corresponding layers. Interestingly, the multisensory units in the intermediate basis function map do not remap all sensory inputs in a common frame of reference, but their visual and auditory receptive fields are partially shifting and gain modulated by eye position, as have been observed in several multisensory brain areas. Moreover, the degree of shifts of these receptive fields with the eye depends on the relative strength of the visual and auditory inputs to the multisensory layer. Thus the model can account for the distribution of shifts observed in experiments as reflecting fluctuations of the ratio between visual and auditory weight among multisensory cells.

We showed that in the presence of noisy sensory and postural inputs these networks converge to the maximum likelihood estimates of their noisy input, which also correspond to the Bayesian estimates for flat prior probabilities (see Fig. 4c). In the particular example we consider, the stable patterns of activities on the visual and auditory layers encode the most likely position of the stimulus given the initially noisy visual and auditory input. To illustrate the Bayesian properties of the network we used it produce results analogs to Banks et al. data [14]. The activities on the eye-centered layer were interpreted as encoding the width of the visual bar, while the activities on the head-centered layer were interpreted as encoded the width of the haptic bar. The multi-directional sensory prediction between visual and haptic responses is not completely equivalent to the coordinate transform implemented by the network, however we showed that our results would generalize to all networks performing multi-directional sensory prediction and, in particular, to a network implementing the real visual to haptic transform. We implemented the visual and haptic sensory noise by adding poisson noise to the visual and haptic inputs clamped in each input layers, and we matched the unimodal discrimination threshold measured by Banks et al. by adjusting the gains of these inputs. We then measured the discrimination thresholds of the network estimates with bimodal visual and haptic inputs, and the weights given to the visual modality in these bimodal estimates. The results are plotted in Fig. 3a and b together with the experimental data and the Bayesian prediction. As expected, the visual weights and the discrimination thresholds of the network estimates are very close to the prediction made by the Bayesian hypothesis, and thus a good match for the experimental data.

This network is designed to implement a spatial link between modalities and perform optimal multisensory integration. Thus, it can also account for cross-modal spatial attention. For example, enhanced activities in a local portion of the visual map (corresponding to attention focused at a particular eye-centered position) will propagate to the auditory layer at the corresponding head-centered location through the multisensory layer, taking into account the eye position. Similarly, if we interpret the head-centered layer as a tactile layer, a tactile stimulus will enhance visual responses both in the multisensory layers and in the visual unimodal layer at the corresponding eye-centered position. Thus, according to this view, direct cross-modal influences on unimodal sensory areas are implemented by feedback connections from the multisensory areas to the unimodal visual areas. This, in turn, account for the existence of both multisensory area and direct cross-modal influences on unimodal brain areas as observed in brain imaging [25,26].

The multi-directional sensory prediction model described above performs Bayesian integration using the gain of the hills of activity to represent the reliability of a sensory input (Fig. 4b) [8]. One limitation of this approach is that it assumes that all sensory inputs come from a unique location in space. As a consequence, the network fail when confronted with several objects, because it integrates visual and auditory inputs whether or not they actually belong to the same object. Clearly, control mechanisms are needed to regulate which sensory inputs should be integrated and how much one sensory modality is allowed to influence the sensory representations in another modality.

Another weakness of the model we have just described is its propensity to amplify weak inputs iteratively (through the feed-forward/feedback loop linking input layers and multisensory layer). This could result in pure sensory noise being interpreted as a phantom location for a non-existent object. Thus, the threshold under which sensory inputs fails to drive the network and be amplified has to be adjusted at a level where the system can be confident an object is really present. This, in itself, is sufficient to account for multisensory enhancement and inverse effectiveness, as observed in the superior colliculus. Bimodal inputs reach the minimum threshold more easily than unimodal inputs, resulting in multisensory enhancement for weak sensory input. This effect disappear for stronger stimuli, when both unimodal and bimodal inputs reach the threshold, hence the inverse effectiveness. At this stage, it is unclear whether multisensory enhancement and inverse effectiveness are due to the fact that SC neurons represent the probability of a target being present (Anastasio et al. [1]), or whether it is due to an iterative amplification of sensory inputs through multi-directional sensory prediction.

6. Conclusion

We have reviewed several recent studies showing that humans can perform Bayesian context-dependent multisensory integration. This process could be implemented with population patterns of activity representing probability distributions over the sensory variables (Fig. 4). However, we saw that this approach runs into two major problems. First, this representational scheme is not realistic in all situations, and in particular when contrast is being manipulated. Second, this approach does not explain how sensory modalities are combined despite using different frames of reference in their early stages.

The alternative to an explicit representation of probability distributions is to use the model proposed by Deneve et al., in which multisensory areas combine sensory and postural inputs in a format allowing multi-directional sensory predictions. This model accounts for neurophysiological and psychophysical data and performs Bayesian multisensory integration without explicitly representing probability distributions.

There are several hypotheses behind the multi-directional sensory prediction model that remain to be explored. First, the model posits a strong link between attention, probability and reliability, all encoded in the gain of the neural responses (and not in changes in the shape of their tuning curves). There are already evidences for links between neural responses gain and attention [28], as well as between neural responses gain and probability [4]. However, it remains to be seen whether neural responses gain scales with the reliability, or equivalently the log likelihood, of the variables they represent. Second, the model proposes that some multisensory interactions could be implemented in part by feedback from multisensory areas to unimodal sensory areas. Direct cross-modal influences, which have been evidenced through brain imaging studies, have not been reported in electrophysiological recordings from single neurons.

Finally, although most studies have focused on spatial representations, it is important to keep in mind that multisensory integration, and more generally, cue integration, extends to many other domains such as depth perception, motion discrimination, speech comprehension, and object identification. Interestingly, many of the concepts we have reviewed here readily generalize other situations, indicating that the Bayesian framework may provide a general theory of perception [22].

Acknowledgements

A.P. and S.D. were supported by a fellowship from the Sloan Foundation, a young investigator award from ONR (N00014-00-1-0642) and a research grant from the McDonnell-Pew foundation.

References

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

[2] R.A. Andersen, G.K. Essick, R.M. Siegel, Encoding of spatial location by posterior parietal neurons, Science 230 (1985) 456–458.

[3] J.E. Atkins, J. Fiser, R.A. Jacobs, Experience-dependent visual cue integration based on consistencies between visual and haptic percepts, Vision Res. 41 (2001) 449–461.

[4] M.A. Basso, R.H. Wurtz, Modulation of neuronal activity in superior colliculus by changes in target probability, J. Neurosci. 18 (1998) 7519–7534.

[5] A. Batista, C. Buneo, L. Snyder, R. Andersen, Reach plans in eye-centered coordinates, Science 285 (1999) 257–260.

[6] A. Blake, H.H. Bulthoff, D. Sheinberg, Shape from texture: ideal observers and human psychophysics, Vision Res. 33 (1993) 1723–1737.

[7] C.A. Buneo, M.R. Jarvis, A.P. Batista, R.A. Andersen, Direct visuomotor transformations for reaching, Nature 416 (2002) 632–636.

[8] S. Deneve, P. Latham, A. Pouget, Efficient computation and cue integration with noisy population codes, Nature Neurosci. 4 (2001) 826–831.

[9] J. Driver, C. Spence, Crossmodal attention, Curr. Opin. Neurobiol. 8 (1998) 245–253.

[10] J. Driver, C. Spence, Cross-modal links in spatial attention, Philos. Trans. R. Soc. Lond. B: Biol. Sci. 353 (1998) 1319–1331.

[11] J. Driver, C. Spence, Multisensory perception: beyond modularity and convergence, Curr. Biol. 10 (2000) R731–R735.

[12] J. Duhamel, F. Bremmer, S. BenHamed, W. Graf, Spatial invariance of visual receptive fields in parietal cortex neurons, Nature 389 (1997) 845–848.

[13] M. Eimer, J. Driver, Crossmodal links in endogenous and exogenous spatial attention: evidence from event-related brain potential studies, Neurosci. Biobehav. Rev. 25 (2001) 497–511.

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

[15] M.O. Ernst, M.S. Banks, H.H. Bulthoff, Touch can change visual slant perception, Nature Neurosci. 3 (2000) 69–73.

[16] M.A. Frens, A.J. Van Opstal, R.F. Van der Willigen, Spatial and temporal factors determine auditory-visual interactions in human saccadic eye movements, Percept. Psychophys. 57 (1995) 802–816.

[17] M. Graziano, X. Hu, C. Gross, Visuospatial properties of ventral premotor cortex, J. Neurophysiol. 77 (1997) 2268–2292.

[18] M.S. Graziano, G.S. Yap, C.G. Gross, Coding of visual space by premotor neurons, Science 266 (1994) 1054–1057.

[19] M.F. Jay, D.L. Sparks, Sensorimotor integration in the primate superior colliculus: I. Motor Convergence, J. Neurophysiol. 57 (1987) 22–34.

[20] S. Kennett, M. Eimer, C. Spence, J. Driver, Tactile-visual links in exogenous spatial attention under different postures: convergent evidence from psychophysics and ERPs, J. Cognitive Neurosci. 13 (2001) 462–478.

[21] D. Knill, Surface orientation from texture: ideal observers, generic observers and the information content of texture cues, Vision Res. 38 (1998) 1655–1682.

[22] D.C. Knill, W. Richards, Perception as Bayesian Inference, Cambridge University Press, Cambridge, MA, 1996.

[23] E. Ladavas, F. Pavani, A. Farne, Auditory peripersonal space in humans: a case of auditory-tactile extinction, Neurocase 7 (2001) 97–103.

[24] E. Ladavas, G.D. Pellegrino, A. Farne, G. Zeloni, Neuropsychological evidence of an integrated visuotactile representation of peripersonal space in humans, Cognitive Neurosci. 10 (1998) 581–589.

[25] E. Macaluso, J. Driver, Spatial attention and crossmodal interactions between vision and touch, Neuropsychologia 39 (2001) 1304–1316.

[26] E. Macaluso, C.D. Frith, J. Driver, Crossmodal spatial influences of touch on extrastriate visual areas take current gaze direction into account, Neuron 34 (2002) 647–658.

[27] J.B. Mattingley, J. Driver, N. Beschin, I.H. Robertson, Attentional competition between modalities: extinction between touch and vision after right hemisphere damage, Neuropsychologia 35 (1997) 867–880.

[28] C.J. McAdams, J.H. Maunsell, Attention to both space and feature modulates neuronal responses in macaque area V4, J. Neurophysiol. 83 (2000) 1751–1755.

[29] M. Mon-Williams, J.P. Wann, M. Jenkinson, K. Rushton, Synaesthesia in the normal limb, Proc. R. Soc. Lond. B: Biol. Sci. 264 (1997) 1007–1010.

[30] B.E. Stein, M.A. Meredith, The Merging of the Senses, MIT Press, Cambridge, MA, 1994.

[31] B.E. Stein, M.A. Meredith, W.S. Huneycutt, L. McDade, Behavioral indices of multisensory integration: orientation to visual cues is affected by auditory stimuli, J. Cognitive Neurosci. 1 (1989) 12–24.

[32] B. Stricanne, R. Andersen, P. Mazzoni, Eye-centered, head-centered, and intermediate coding of remembered sound locations in area LIP, J. Neurophysiol. 76 (1996) 2071–2076.

[33] R.J. van Beers, A.C. Sittig, J.J. Denier van der Gon, How humans combine simultaneous proprioceptive and visual position information, Exp. Brain Res. 111 (1996) 253–261.

[34] R.J. van Beers, A.C. Sittig, J.J. Denier van der Gon, The precision of proprioceptive position sense, Exp. Brain Res. 122 (1998) 367–377.

[35] R.J. van Beers, A.C. Sittig, J.J. Gon, Integration of proprioceptive and visual position-information: An experimentally supported model, J. Neurophysiol. 81 (1999) 1355–1364.

[36] R.J. van Beers, D.M. Wolpert, P. Haggard, When feeling is more important than seeing in sensorimotor adaptation, Curr. Biol. 12 (2002) 834–837.

[37] M.T. Wallace, M.A. Meredith, B.E. Stein, Multisensory integration in the superior colliculus of the alert cat, J. Neurophysiol. 80 (1998) 1006–1010.

[38] M.T. Wallace, L.K. Wilkinson, B.E. Stein, Representation and integration of multiple sensory inputs in primate superior colliculus, J. Neurophysiol. 76 (1996) 1246–1266.

[39] L.M. Ward, J.J. McDonald, D. Lin, On asymmetries in cross-modal spatial attention orienting, Percept. Psychophys. 62 (2000) 1258–1264.

[40] D. Warren, H. Pick, Intermodality relations in localization in blind and sighted people, Percept. Psychophys. 8 (1970) 430–432.

[41] M.J. Young, M.S. Landy, L.T. Maloney, A perturbation analysis of depth perception from combinations of texture and motion cues, Vision Res. 33 (1993) 2685–2696.

[42] R.S. Zemel, P. Dayan, A. Pouget, Probabilistic interpretation of population codes, Neural Comput. 10 (1998) 403–430.