Sound-induced flash illusion as an optimal percept

Ladan Shams[a,b], Wei Ji Ma[b] and Ulrik Beierholm[c]

[a]Department of Psychology, University of California Los Angeles, Los Angeles, [b]Division of Biology and [c]Computation and Neural Systems, California Institute of Technology, Pasadena, California, USA.

Correspondence and requests for reprints to Ladan Shams, PhD, Department of Psychology, University of California Los Angeles, Franz Hall 7445B, Los Angeles, CA 90095-1563, USA
E-mail: ladan@psych.ucla.edu

Sponsorship: W.J.M. was supported by the Swartz Foundation and the Netherlands Organization for Scientific Research. U.B. was supported by the David and Lucile Packard Foundation.

Received 3 August 2005; accepted 21 September 2005

Recently, it has been shown that visual perception can be radically altered by signals of other modalities. For example, when a single flash is accompanied by multiple auditory beeps, it is often perceived as multiple flashes. This effect is known as the sound-induced flash illusion. In order to investigate the principles underlying this illusion, we developed an ideal observer (derived using Bayes' rule), and compared human judgements with those of the ideal observer for this task. The human observer's performance was highly consistent with that of the ideal observer in all conditions ranging from no interaction, to partial integration, to complete integration, suggesting that the rule used by the nervous system to decide when and how to combine auditory and visual signals is statistically optimal. Our findings show that the sound-induced flash illusion is an epiphenomenon of this general, statistically optimal strategy. *NeuroReport* 16:1923–1927 © 2005 Lippincott Williams & Wilkins.

Keywords: auditory–visual perception, Bayesian inference, cross-modal illusion, cue combination, ideal observer, multisensory integration, multisensory perception, sound-induced flash illusion

Introduction

Situations in which an individual is exposed to sensory signals in only one modality are the exception rather than the rule. At any given instant, the brain is typically engaged in processing sensory stimuli from two or more modalities, and in order to achieve a coherent and ecologically valid perception of the physical world, it must determine which of these temporally coincident sensory signals are caused by the same physical source/event and thus should be integrated into a single percept. Spatial coincidence of the stimuli is not a very strong or exclusive determinant of cross-modal binding: information from two modalities may get seamlessly bound together, despite large spatial inconsistencies (e.g. ventriloquism effect), while spatially concordant stimuli may be perceived as separate entities (e.g. someone speaking behind a screen does not lead to the binding of the voice with the screen). This is not surprising, considering the relatively poor spatial resolution of auditory, olfactory, and somatosensory modalities. The degree of consistency between the information conveyed by two sensory signals, on the other hand, is clearly an important factor in determining whether the cross-modal signals are to be integrated or segregated.

Previous models of cue combination [1–9] have all focused exclusively on conditions in which the signals of the different modalities get completely integrated (or appear so because of the employed paradigms that force participants to report only one percept, and thus not revealing any potential conflict in percepts). Therefore, the previous models are unable to account for the vast number of situations in which the signals do not get integrated or only partially integrate.

The sound-induced flash illusion [10,11] is a psychophysical paradigm in which both integration and segregation of auditory–visual signals occur depending on the stimulus condition. When one flash is accompanied by one beep (i.e. when there is no discrepancy between the signals), the single flash and single beep appear to originate from the same source, and are completely fused. When one flash is accompanied by four beeps (i.e. when the discrepancy is large), however, most often they are perceived as emanating from two separate events, and the two signals are segregated, that is, a single flash and four beeps are perceived. If the single flash is accompanied by two beeps (i.e. when the discrepancy is small), the single flash is often perceived as two flashes and on these illusion trials, the flashes and beeps are perceived as having originated from the same source, that is, integration occurs in a large fraction of trials. When a single flash is accompanied by three beeps, on a fraction of trials the single flash is perceived as two flashes while the three beeps are perceived as veridical. These trials would exemplify conditions of partial integration in which the visual and/or auditory percepts are shifted towards each other, but do not converge.

Therefore, the sound-induced flash illusion offers a paradigm encompassing the entire spectrum of bisensory situations. As signals are not always completely integrated, previous models of cross-modal integration cannot account

for these effects. Therefore, we developed a new model in order to be able to account for situations of segregation and partial integration, as well as complete integration. The model is an ideal observer and in contrast to previous models of cue combination, it does not assume one source for all the sensory signals (which would enforce integration); instead, it assumes one source for the signal in each modality. The sources, however, are not taken to be statistically independent, and therefore, the model allows inferences about both cases in which separate entities have caused the sensory signals, and cases in which sensory signals are caused by one source. The model uses Bayes' rule to make inferences about the causes of the various sensory signals.

We presented observers with varying combinations of beeps and flashes, and asked them to report the perceived number of flashes and beeps in each trial. We then compared the human judgements with those of the ideal observer.

Materials and methods

Stimuli

The visual stimulus consisted of a uniform white disk subtending 1.5° of the visual field at 12° eccentricity below the fixation point (Fig. 1a), flashed for 10 ms on a black computer screen 1–4 times. The auditory stimulus was a 10-ms-long beep with 80 dB sound pressure level and 3.5 kHz frequency, also presented 0–4 times. A factorial design was used in which all combinations of 0–4 flashes and 0–4 beeps (except for the no flash–no beep combination) were presented, leading to a total of 24 conditions. The stimulus onset asynchronies (SOAs) of flashes and beeps were 70 and 58 ms, respectively (Fig. 1b). These specific SOAs were chosen because of certain constraints (e.g. frame rate, obtaining a strong illusion in the illusion conditions, and the smallest sound SOA, which consistently is above flutter fusion threshold). The behavioral data are fairly robust to the exact visual and auditory SOAs. The relative timing of the flashes and beeps was set such that the centers of the flash and beep sequences were synchronous in order to maximize the time overlap between the two stimuli. Sound was presented from two speakers placed adjacent to the two sides of the computer monitor, at the height in which the visual stimulus was presented, thus, localizing at the same location as the visual stimulus.

Procedure

Ten naive observers participated in the experiment. Observers sat at a viewing distance of 57 cm from the computer screen and speakers. Throughout the trials, there was a constant fixation point at the center of the screen. The observer's task was to judge both the number of flashes seen and the number of beeps heard after each trial (these reports provide $P(Z_A,Z_V \mid A,V)$ as described below). The experiment consisted of 20 trials of each condition, amounting to a total of 480 trials, ordered randomly. A brief rest interval was given after every third trial of the experiment.

The ideal observer model

We assume that the auditory and visual signals are statistically independent given the auditory and visual causes (see Fig. 2). This is a common assumption, motivated by the hypothesis that the noise processes that corrupt the auditory and visual signals are independent. This conditional independence means that if the causes are known, knowledge about V provides no information about A, and vice versa, as the noises corrupting the two signals are independent. In the meantime, if the causes are not known, knowledge of V provides information about A, and vice versa [12].

The information about the likelihood of sensory signal A occurring, given an auditory cause Z_A, is captured by the probability distribution $P(A \mid Z_A)$. Similarly, $P(V \mid Z_V)$ represents the likelihood of sensory signal V given a source Z_V in the physical world. The priors $P(Z_A,Z_V)$ denote the perceptual knowledge of the observer about the auditory–visual events in the environment. In addition to the observer's experience, the priors may also reflect hard-wired biases imposed by the physiology and anatomy of the brain (e.g. the pattern of interconnectivity between the sensory areas [13,14]), as well as biases imposed by the task, the observer's state, etc.

The graph in Fig. 2 [15] illustrates the two key features of the model. First, that there are two sources, Z_A and Z_V, for the two sensory signals A and V. This allows inference in both cases in which the signals A and V are caused by the same source and cases in which they are caused by two distinct sources. That is, in contrast to the previous models, this model does not *a priori* assume that the signals have to be integrated. Second, in this model, Z_V influences A only through its effect on Z_A, and likewise for Z_A and V. This corresponds to the assumption of independent likelihood functions, $P(A,V \mid Z_A,Z_V)=P(A \mid Z_A)P(V \mid Z_V)$. This is a plausible assumption motivated by the fact that either the two signals are caused by two different events in which case A would be independent of Z_V (and likewise for V and Z_A), or they are caused by one event, in which case the dependence of A on Z_V can be captured by its dependence on Z_A.

Given the visual and auditory signals A and V, an ideal observer would try to make the best possible estimate of the physical sources Z_A and Z_V, based on the knowledge $P(A \mid Z_A)$, $P(V \mid Z_V)$, and $P(Z_A,Z_V)$. These estimates are based on the posterior probabilities $P(Z_A,Z_V \mid A,V)$, which can be calculated using Bayes' rule, and simplified by the assumptions represented by the model structure (Fig. 2), resulting in the following inference rule:

$$P(Z_A, Z_V|\mathrm{A}, \mathrm{V}) = \frac{\mathrm{P(A|Z_A)\ P(V|Z_V)\ P(Z_A, Z_V)}}{\mathrm{P(A, V)}}. \qquad (1)$$

This inference rule simply states that the posterior probability of events Z_A and Z_V is the normalized product of the single-modality likelihoods and joint priors. In order to simplify calculations, we assume that $P(A,V)$ has a uniform distribution. This, in turn, implies that $P(A)$ and $P(V)$ also have uniform distributions. Given a uniform $P(A)$, the auditory likelihood term is computed as follows

$$P(A|\mathrm{Z_A}) = \frac{\mathrm{P(Z_A|A)P(A)}}{\sum_{\mathrm{A}} \mathrm{P(Z_A|A)P(A)}} = \frac{\mathrm{P(Z_A|A)}}{\sum_{\mathrm{A}} \mathrm{P(Z_A|A)}}$$

(and likewise for $P(V \mid Z_V)$). While the likelihood functions $P(A \mid Z_A)$ and $P(V \mid Z_V)$ are nicely approximated from the unisensory (visual-alone and auditory-alone) conditions, the prior probabilities $P(Z_A,Z_V)$ involve both sensory modalities and cannot be obtained from unisensory conditions alone.

Estimation of the joint priors

In most models, the priors are not directly computable. Hence, the prior distribution is parameterized and the

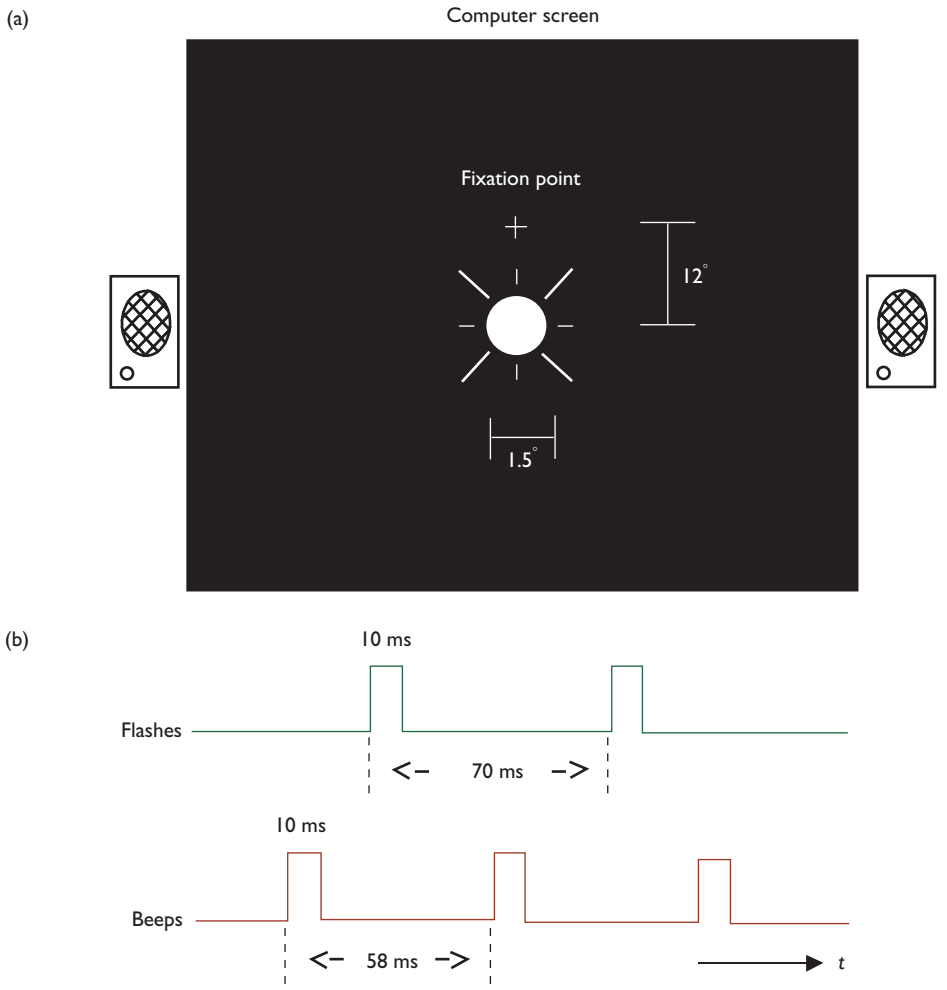

Fig. 1 The spatio-temporal configuration of stimuli. (a) The spatial configuration of the stimuli. The visual stimulus was presented at 12° eccentricity below the fixation point. The sounds were presented from the speakers adjacent to the monitor and at the same height as the center of the visual stimulus. (b) The temporal profile of the stimuli in one of the conditions (2 flashes + 3 beeps) is shown. The centers of the visual and auditory sequences were aligned in all conditions.

parameters are tuned to fit the observed data (i.e. data to be predicted). Our experimental paradigm makes it possible for the joint priors to be approximated directly from the observed data, alleviating the need for any parameter tuning. The joint priors can be approximated by marginalizing the joint probabilities across all conditions, that is, all combinations of A and V:

$$P(Z_A, Z_V) = \sum_{A,V} P(Z_A, Z_V|\mathrm{A}, \mathrm{V})\ P(\mathrm{A}, \mathrm{V}). \qquad (2)$$

Given a uniform $P(A,V)$, this leads to a normalized marginalization of the posteriors. As this estimate requires marginalizing over all conditions including auditory–visual conditions, we used the data from a different set of observers (the first half of participants) for estimating the joint priors using the above formula, and excluded those data from the testing process (the second half of participants). In other words, these data were used only for calculating the priors and discarded afterwards. Thus, the model remained predictive, not using any auditory–visual data for making predictions about the performance in the auditory–visual conditions.

Although it may appear that the joint prior matrix introduces 24 free parameters in our model, it should be emphasized that this is not the case, as these parameters are not 'free'. The parameters of the joint prior matrix are set using the observed data; however, they were not tuned to minimize the error between the model predictions and the data. Therefore, the model has no 'free' parameters.

Results

The observers perform better in the auditory-alone conditions (first row of Fig. 3) than in the visual-alone conditions (first column of Fig. 3). As can be seen in Fig. 3, the human observer's performance is remarkably consistent with that of the ideal observer in all of the conditions (r^2=0.92), accounting for 600 data points [(25 (Z_A,Z_V) combinations at 24 conditions] with no free parameters.

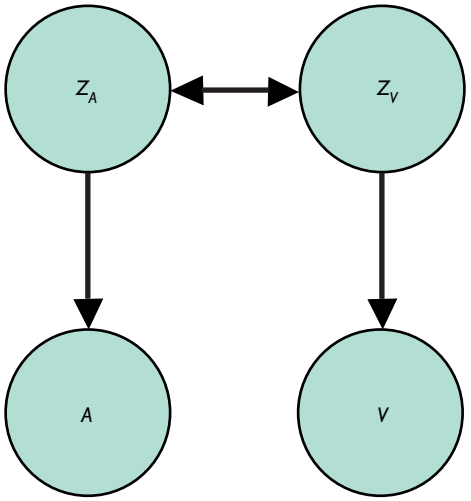

Fig. 2 Graphical model describing the ideal observer. In a graphical model [15], the graph nodes represent random variables, and arrows denote potential conditionality. The absence of an arrow represents direct statistical independence between the two variables. The bidirectional arrow between Z_A and Z_V does not imply a recurrent relationship; it implies that the two causes are not necessarily independent.

Only in conditions in which the visual and auditory stimuli are identical (i.e. the conditions displayed along the diagonal) do observers consistently indicate perceiving the same number of events in both modalities. In conditions in which the inconsistency between the auditory and visual stimuli is not too large, for instance, in the 1 flash + 2 beeps condition or 2 flashes + 1 beep condition, there is a strong tendency to combine the two modalities, as indicated by highly overlapping auditory and visual reports. The high values along the diagonal in joint posterior matrices of these conditions (not shown here) confirm that indeed the same number of events were experienced jointly in both modalities, in these conditions. The integration of the auditory–visual percepts is achieved in these cases by a shift of the visual percept in the direction of the auditory percept. This occurs because the variance in the auditory-alone conditions is lower than that of the visual-alone conditions. In other words, because the auditory modality is more reliable, it dominates the overall percept in these auditory–visual conditions. This finding is consistent with previous studies of cue combination within [3,4,16,17] or across modalities [7–9] in all of which the discrepancy between the two cues is small, and the percept is dominated by the cue with lower variance (or higher reliability). The large fraction of trials in

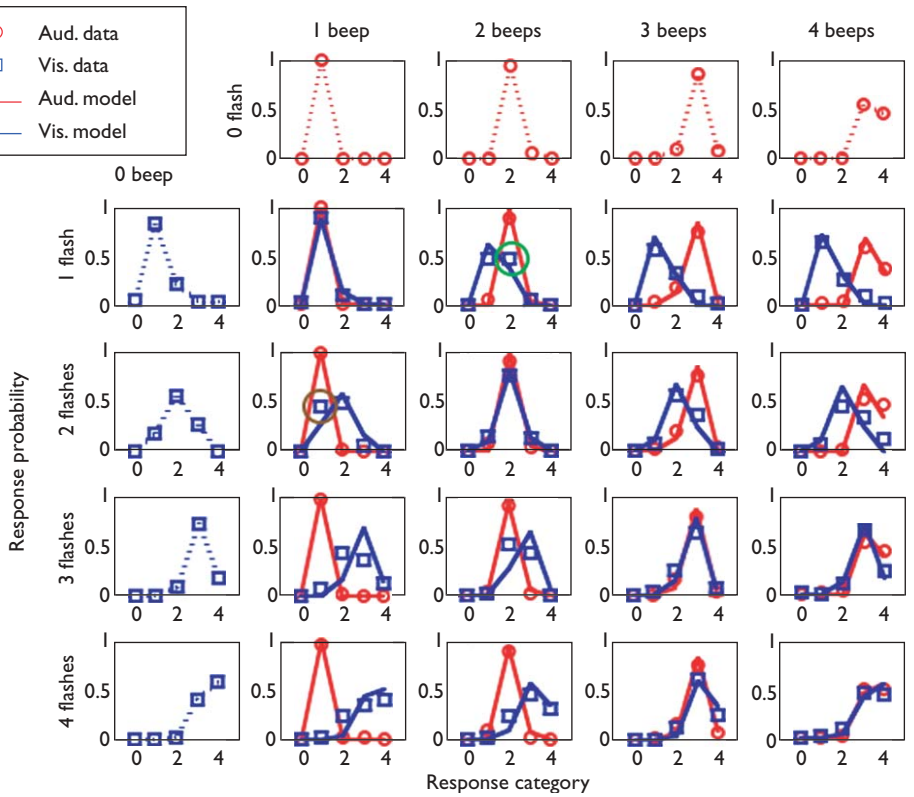

Fig. 3 Comparison of the performance of human observers with the ideal observer. To facilitate interpretation of the data, instead of presenting joint posterior probabilities for each condition, only the marginalized posteriors are shown. The auditory and visual judgements of human observers are plotted in red circles and blue squares, respectively. Each panel represents one of the conditions. The first row and first columns represent the auditory-alone and visual-alone conditions, respectively. The remaining panels correspond to conditions in which auditory and visual stimuli were presented simultaneously. The horizontal axes represent the response category (with zeros denoting absence of a stimulus and 1–4 representing number of flashes or beeps). The vertical axes represent the probability of a perceived number of flashes or beeps. The data point, which is enclosed by a green circle, is an example of the sound-induced flash illusion, showing that in a large fraction of trials, observers perceived two flashes when one flash was paired with two beeps. The data point enclosed by a brown circle reveals an opposite illusion in which two flashes are perceived as one flash in a large fraction of trials in the 2 flashes + 1 beep condition.

which the observers report seeing two flashes in the 1 flash + 2 beeps condition corresponds to the sound-induced flash illusion [10].

In conditions in which the discrepancy between the number of flashes and beeps is large (e.g. 1 flash + 4 beeps or 4 flashes + 1 beep), the overlap between the auditory and visual percepts is significantly smaller, indicating a considerably smaller degree of integration and larger degree of segregation.

Next, we investigated the possibility that the Bayesian model of Eq. (1) is overly powerful and capable of predicting any data set. We shuffled the obtained human observer posterior probabilities $P(Z_A,Z_V \mid A,V)$ in each auditory–visual condition leading to a new data set that was identical to the human data in its overall content, although randomized in order. We applied our model to this data set. The model predictions did not match the shuffled data, even when we did not divide the data set into two halves, and instead computed the priors from the same set for which we generated the predictions ($r^2=-0.05$). We obtain qualitatively similar results regardless of the made-up distribution used. This finding strongly suggests that the predictions of the proposed ideal observer are distinctly consistent with the human observer's data and not with any arbitrary data set.

Discussion

Altogether, these results suggest that humans combine auditory and visual information in an optimal fashion. Our results extend earlier findings (e.g. [7,18]) by showing that the optimality of the human performance is not restricted to situations in which the discrepancy between the two modalities is minute and the two modalities are completely integrated. Indeed, it can be shown that many earlier models of cue combination are special cases of the model described here.

The ideal observer model presented here differs from previous models of cue combination, which have employed maximum likelihood estimation in two important ways. First, as opposed to previous models (which assume one cause for all signals), our model allows a distinct cause for each signal. This is a structural difference between the present model and all the previous models, and is the reason why the multisensory paradigm in the present study is beyond the scope of previous models. The assumption of a single cause makes previous models unable to account for a vast portion of the present data in which the visual and auditory information are not integrated, that is, all the trials in which participants reported different visual and auditory percepts. Second, the previous models did not include any prior probability of events, which is equivalent to assuming a uniform prior distribution. In the present model, prior probabilities are not assumed to be uniform. In order to examine the importance of the priors in accounting for the data, we tested the model using a uniform prior. The goodness of fit was considerably reduced ($r^2=0.62$), indicating that in this task the priors depart significantly from uniform distribution, and are therefore necessary for accounting for the data.

Conclusion

The findings of this study suggest that the brain uses a mechanism similar to Bayesian inference [19] to decide whether, to what degree, and how (in which direction) to integrate the signals from auditory and visual modalities, and that the sound-induced flash illusion can be viewed as an epiphenomenon of a statistically optimal computational strategy.

Acknowledgement

We are grateful to Graeme Smith for extensive discussions and help with programming. We thank Stefan Schaal, Bosco Tjan, Alan Yuille, and Zili Liu for their insightful discussions and comments.

References

1. Bülthoff HH, Mallot HA. Integration of depth modules: stereo and shading. *J Opt Soc Am* 1988; **5**:1749–1758.
2. Knill DC. Mixture models and the probabilistic structure of depth cues. *Vision Res* 2003; **43**:831–854.
3. Landy MS, Maloney LT, Johnston EB, Young M. Measurement and modeling of depth cue combination: in defense of weak fusion. *Vision Res* 1995; **35**:389–412.
4. Yuille AL, Bülthoff HH. Bayesian deicision theory and psychophysics. In: Knill DC, Richards W, editors. *Perception as Bayesian inference*. Cambridge: Cambridge University Press; 1996. pp. 123–161.
5. Alais D, Burr D. The ventriloquist effect results from near-optimal bimodal integration. *Curr Biol* 2004; **14**:257–262.
6. Massaro DW. *Perceiving talking faces: from speech perception to a behavioral principle*. Cambridge, Massachusetts: MIT Press; 1998.
7. Ernst MO, Banks MS. Humans integrate visual and haptic information in a statistically optimal fashion. *Nature* 2002; **415**:429–433.
8. van Beers RJ, Sittig AC, Denier van der Gon JJ. Integration of proprioceptive and visual position information: an experimentally supported model. *J Neurophysiol* 1999; **81**:1355–1364.
9. Ghahramani Z, Wolpert DM, Jordan MI. Computational models of sensorimotor integration. In: Morasso PG, Sanguineti V, editors. *Self-organization, computational maps, and motor control*. Amsterdam: North-Holland, Elsevier Press; 1997. pp. 117–147.
10. Shams L, Kamitani Y, Shimojo S. What you see is what you hear. *Nature* 2000; **408**:788.
11. Shams L, Kamitani Y, Shimojo S. Visual illusion induced by sound. *Cogn Brain Res* 2002; **14**:147–152.
12. Jordan MI. Graphical models. *Stat Sci (Special Issue on Bayesian Stat)* 2004; **19**:140–155.
13. Falchier A, Clavagnier S, Barone P, Kennedy H. Anatomical evidence of multimodal integration in primate striate cortex. *J Neurosci* 2002; **22**: 5749–5759.
14. Rockland KS, Ojima H. Multisensory convergence in calcarine visual areas in macaque monkey. *Int J Psychophysiology* 2003; **50**:19–26.
15. Pearl J. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. San Mateo, California: Morgan Kaufmann, 1988.
16. Clark JJ, Yuille AL. *Data fusion for sensory information processing systems*. Boston: Kluwer Academic Pulishers; 1990.
17. Jacobs R. Optimal integration of texture and motion cues to depth. *Vision Res* 1999; **39**:3621–3639.
18. Battaglia PW, Jacobs RA, Aslin RN. Bayesian integration of visual and auditory signals for spatial localization. *J Opt Soc Am* 2003; **20**: 1391–1397.
19. Knill DC, Pouget A. The Bayesian brain: the role of uncertainty in neural coding and computation. *Trends Neurosci* 2004; **27**:712–719.