

Jan W. H. Schnupp · Karen L. Dawe ·
Gabriella L. Pollack

The detection of multisensory stimuli in an orthogonal sensory space

Received: 11 February 2004 / Accepted: 5 October 2004 / Published online: 15 December 2004
© Springer-Verlag 2004

Abstract The detection of a stimulus can be considerably facilitated if the stimulus engages two or more sensory modalities simultaneously. This phenomenon, commonly referred to as multisensory (or cross-modal) facilitation, has been demonstrated behaviorally in cats and humans. A number of rules are thought to govern this phenomenon. These rules state that strong facilitation is to be expected only if the two sensory modalities are stimulated simultaneously and at the same place, and if the stimuli themselves are weak. However, these rules are not sufficient to allow accurate predictions of multimodal stimulus detection probabilities directly from physical stimulus parameters. Here we show that such predictions are possible on the basis of a simple and biologically plausible psychophysical model, which relates the detection of audio-visual, audio-tactile or visual-tactile stimuli to the Euclidean distance that these stimuli span in an orthogonal sensory space.

Keywords Cross-modal facilitation · Modeling · Multimodal enhancement · Multisensory integration · Psychophysics

Abbreviations D : Deviance · L : Likelihood · N : Number of trials · $N(a, b)$: Normal (Gaussian) distribution with mean a and variance b · S : Neural population signal (thought of as random variable) · $\chi^2(a, b)$: Cumulative chi-Square distribution at a with b degrees of freedom · Δx , Δy : Change in stimulus intensity in stimulus modality x , y · λ : Decision threshold for the detection of a stimulus · μ : Mean of the neural signal · $\Phi(z)$: Cumulative standard normal distribution at z · σ^2 : Variance of the neural signal

Introduction

In the experimental neurosciences, sensory systems are usually investigated by observing the “responses” that are elicited by “stimuli” presented “in isolation”. This reductionist approach has been highly successful, even though it is somewhat artificial, as most organisms evolve in a very rich environment in which numerous stimuli across all modalities compete for our attention. Many “natural” objects or events will stimulate several sensory pathways simultaneously, i.e. they can be seen as well as felt or heard or smelled. It is tempting to assume that our sensory systems have evolved means to exploit the resulting correlations across sensory streams to facilitate sensory processing, and recordings from multisensory (visual-auditory, visual-tactile or auditory-tactile) neurons in the superior colliculus (Meredith and Stein 1983; King and Palmer 1985) gave an early indication that this may indeed be the case. One “rule” that emerged from these studies is that responses of these multisensory neurons could be “enhanced” (i.e. they would exhibit a much larger firing rate) if, and only if, the two sensory modalities were stimulated more or less at the same time and at the same place. Multisensory stimuli emitted by a single real world object usually emanate from a very limited set of positions in space and time (rarely will we see a bell shaking on our left only to hear it ringing 5 min later to our right). Consequently one can easily imagine that multisensory enhancement of neural firing rates which is contingent on temporally and spatially coincident stimulation is designed to facilitate the detection of objects, as it combines evidence across sensory streams in a sensible way. Later psychophysical experiments (Stein et al. 1988; Frassinetti et al. 2002; Lovelace et al. 2003) have lent additional support to this interpretation.

Stein and Meredith (1993) summarized their observations on cross-modal integration by formulating a set of “rules” of multisensory enhancement. These state that the magnitude to the enhancement depends on the degree of spatial and temporal coincidence across sensory modalities, but also include the “inverse effectiveness” rule,

J. W. H. Schnupp (✉) · K. L. Dawe · G. L. Pollack
University Laboratory of Physiology, University of Oxford,
Parks Road,
Oxford, OX1 3PT, UK
e-mail: jan.schnupp@physiol.ox.ac.uk
Tel.: +44-1865-272513

which states, “maximal enhancement occurs with minimally effective stimuli”. The inverse effectiveness rule encapsulates the notion that cross-modal enhancement is most useful if, in a bimodal stimulus, neither of the individual modality stimuli are intense enough to guarantee detection. Conversely, if either stimulus component (or both) is already a highly effective stimulus on its own, then there is little need for cross-modal enhancement.

These rules of multisensory enhancement are semi-quantitative, in that they make some predictions about the relative expected magnitude of enhancement, but on their own they are not sufficient to allow us to predict response magnitudes accurately and directly from knowledge of stimulus parameters alone. A quantitative model capable of completely encapsulating all of Stein and Meredith’s rules of multisensory integration should be able to make accurate predictions of detection probabilities on the basis of three sets of variables, namely the respective intensities, spatial positions and relative time of occurrence of the unimodal stimulus components. In this paper we take a first step towards such a quantitative re-formulation by re-examining the dependence of detection probability on the respective unimodal intensities for stimuli that are presented synchronously and from a single location.

Results

In our search for a quantitative model, our starting point was the notion that sensory systems have evolved to be much more sensitive to changes in the current sensory scene than to absolute stimulus intensities. We constructed a small device, to be held by the experimental subject, which would deliver stimuli in the form of small, brief increments in the intensity of continuously ongoing acoustic, visual or vibro-tactile stimulation. In the spirit of Weber’s law, which states that the smallest detectable change in the intensity of a sensory stimulus is a constant fraction of the original (background) stimulus magnitude, we quantified all stimulus levels as fractional increments relative to the background intensity. Further details are given in the methods section. Quantitative models capable of predicting the observed multisensory stimulus detection rates were then developed from within the framework of signal detection theory (Wickens 2002). At the heart of this approach lies the assumption that stimuli are represented in the brain by a “neural signal” S . The precise nature of this signal is still unknown, but it seems highly likely that it is carried by a population code, possibly a “population vector” (Georgopoulos et al. 1986) of activity, perhaps in the superior colliculus (King and Schnupp 1999) or a multisensory cortical area (Wallace et al. 1992). Given the highly stochastic nature of neural responses, the signal S must be thought of as a random

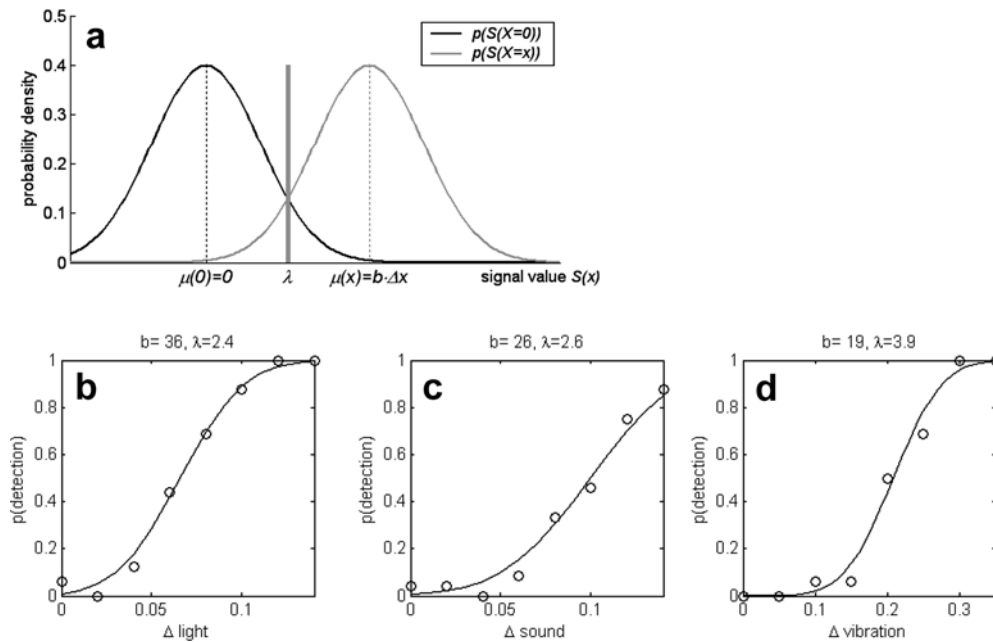


Fig. 1 **a** Threshold signal detection model. It is assumed that the neural population signal that represents ongoing background stimulus levels is a random variable S with a Gaussian distribution (black curve). After suitable normalization, this distribution can be assumed to have zero mean and unit variance. The effect of presenting a stimulus is to shift the mean of the distribution from zero to $b \cdot \Delta x$ (gray curve), where Δx denotes the relative increase in stimulus level. Stimulus detection is therefore equivalent to deciding whether a “random sample” taken from the neural signal S at the moment of decision comes from the black or the gray distribution.

This is thought to be achieved by comparing the observed signal value to a decision boundary λ . **b–d** Probability of stimulus detection as a function of stimulus level for unimodal cases. The threshold detection model illustrated in **a** predicts that detection probabilities should conform to a cumulative Gaussian distribution ($p_{\lambda} = \Phi(b \cdot \Delta x - \lambda)$). The open circles show observed data points. The data shown in **b–d** respectively are representative examples taken from three different subjects. The solid lines show the cumulative Gaussians fitted to the data by estimating b and λ using maximum likelihood methods

variable. The probability distributions governing individual neural discharge patterns are also uncertain, although it is widely assumed that they are Poisson distributed (Colonius and Diederich 2001). However, if S is indeed carried by a population code, then, due to the central limit theorem, its probability distribution will be Gaussian regardless of the statistical nature of the responses of individual neurons. Both the mean μ and the variance σ^2 of this Gaussian distribution are not known a priori and may depend on stimulus level. We make the simple assumptions that the variance is a constant independent of the stimulus and that the mean grows in proportion to the size of the stimulus level, i.e. $\mu(x)=\mu(0)+b\cdot\Delta x$. Thus, b determines the sensitivity of the observer to the stimulus: if b is large, then even weak stimuli (small Δx) can produce relatively large changes in the average neural signal $\mu(x)$.

Conveniently, it should always be possible to measure the neural signal S in suitably normalized units, such that the mean for the background intensity $\mu(0)$ is zero and the value of σ^2 is one. After normalization, the neural signal in the absence of a stimulus ($S(X=0)$) will have a standard normal distribution ($S(X=0) \sim N(0,1)$) while the signal evoked by a stimulus of level x has a distribution shifted proportionally to x ($S(X=x) \sim N(b\cdot\Delta x,1)$). (Readers versed in psychophysical theory may note that the value $b\cdot\Delta x$ would equal the d' sensitivity index commonly used in signal detection theory). As illustrated in Fig. 1a, the effect of presenting a stimulus of level x is therefore to shift the distribution of S from its “background” mean $\mu(0)=0$ to the new value $\mu(x)=b\cdot\Delta x$. To decide whether a stimulus was indeed present, the brain of the subject is thought to compare the value of S to a decision criterion (or “threshold”) λ . The observer reports the detection of a stimulus when the threshold is exceeded ($S>\lambda$). Consequently, this simple detection model predicts that the probability of stimulus detection should be given by

$$p_x = \Phi(b \cdot \Delta x - \lambda) \quad (1)$$

where $p_x = p(D|X=\Delta x)$ is the probability that the subject reported the detection of a stimulus given that an intensity step of size Δx was delivered in modality X , and Φ is the cumulative standard normal distribution. Like b , the value of the criterion λ for any particular observer is not known a priori, but can be estimated from the data. Here we used a gradient descent—maximum likelihood method (see Methods) to estimate the values of b and λ . In all cases, the threshold detection model (Eq. 1) produced a good fit of the observed unimodal detection probabilities. Representative examples are shown in Fig. 1b–d. Weak stimuli shift $\mu(x)$ by only a small amount, resulting in considerable overlap in the distribution of signals experienced in the absence or the presence of a stimulus. In that case it is not possible to place λ so as to allow error-free detection of a stimulus. The observer will make “false alarms”, i.e. report a stimulus where there was none, when $S(X=0)$ falls in the upper tail of its distribution so that $S(X=0)>\lambda$.

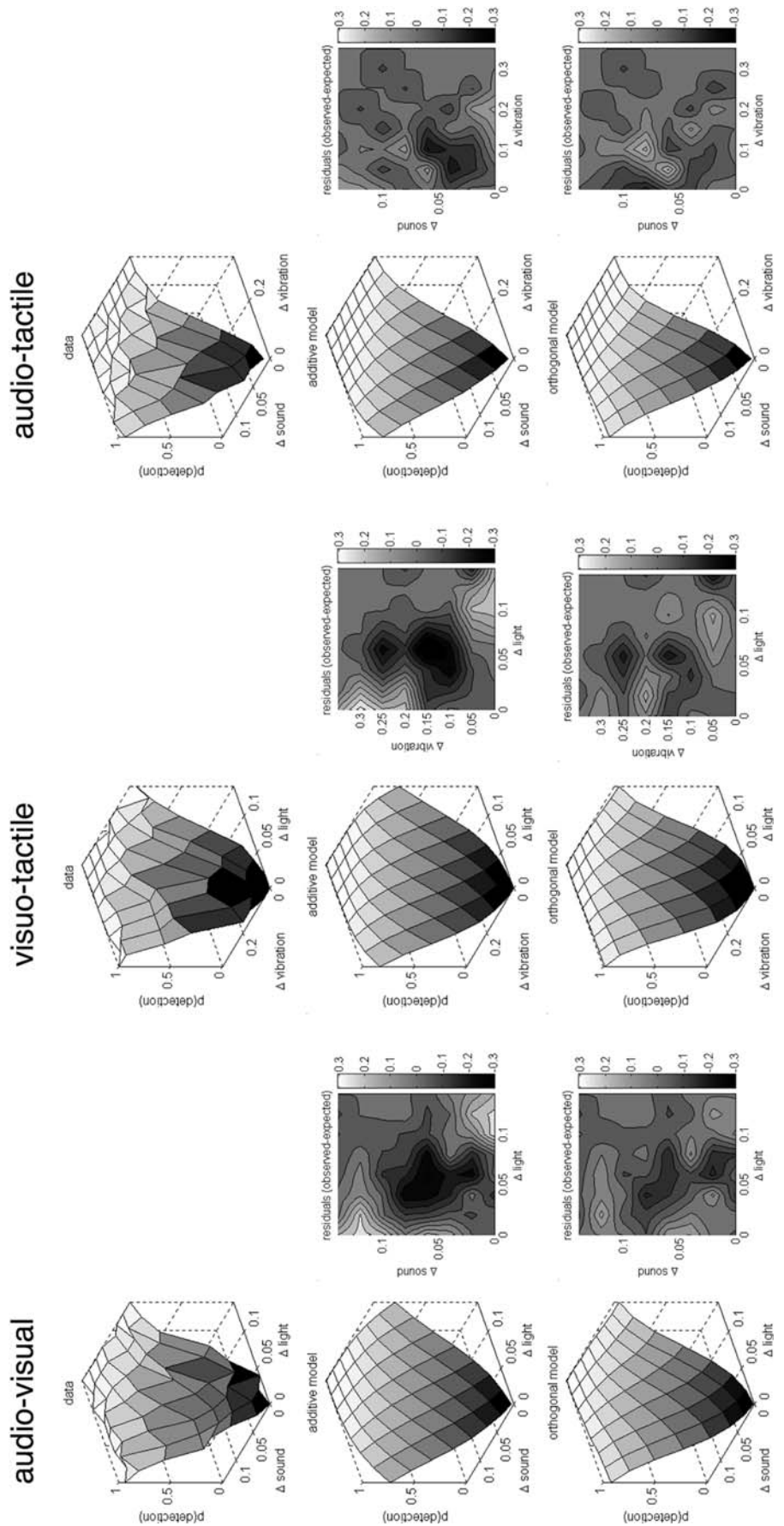
Similarly, the observer will fail to detect a stimulus whenever $S(X=x)<\lambda$. Large values of λ result in small false alarm rates but frequent misses, while small values of λ have the opposite effect. The simple detection model (Eq. 1) observed values for λ varied somewhat from subject to subject (mean 2.06, std 0.65), which corresponds to false alarm rates of, on average, around 1%. For b the values depended to some extent on stimulus modality. Subjects were typically more sensitive to the visual (mean \pm std for $b=31.8\pm 5.3$) than the auditory ($b=21.4\pm 6.0$) or vibro-tactile ($b=12.3\pm 3.3$) intensity steps.

For unimodal stimuli, our data clearly conform with predictions derived from a well-established elementary signal detection model. The crucial question for this paper is how this model should be extended to account for the detection of bimodal stimuli. In that context, there are a number of possible hypotheses to consider. One possibility considered by King and Schnupp (1999) is that the neural signals S_x, S_y associated with stimulus modalities X, Y might simply be added in the brain. In terms of our simple detection model (Eq. 1) the additivity hypothesis implies that, if stimulus Δx in modality x will on average increase S by $b_x\cdot\Delta x$ and stimulus Δy in modality y will increase S by $b_y\cdot\Delta y$, then the combined average effect of x and y should be $\mu(x,y)=b_x\cdot\Delta x+b_y\cdot\Delta y$, and Eq. 1 extends to

$$p_{xy} = \Phi(b_x \cdot \Delta x + b_y \cdot \Delta y - \lambda) \quad (2)$$

However, this attractively simple additive model was unable to provide an adequate fit to most of our data. In 12 out of 17 cases the deviances of the best fit additive models were so large that the models could be rejected at the 5% level, and for six out of 17 cases the models could be rejected at $p<0.0001$. Representative examples of data and corresponding model fits are shown in Fig. 2. The *first* row of panels in Fig. 2 shows raw data for audio-visual, visuo-tactile and audio-tactile stimulus combinations respectively. The second row shows the corresponding maximum likelihood fits for the additive model. Residuals (differences between observed and predicted values) are also shown as interpolated contour plots. For all three modality combinations, the residual plots reveal systematic deviations between model and data, notably relatively large negative residuals around the main diagonal of the plot, flanked by predominantly positive residuals along the left and bottom margins. This, as well as a direct comparison of the raw data and the model predictions, suggests that the additive model fails to account for the bimodal data because it produces a function that is “not concave enough”, i.e. it systematically overestimates the detection rates for weak bimodal stimuli while underestimating that of unimodal ones. This observation leads to an important conclusion: the “signals” thought to represent the individual modalities within the observer’s brain clearly interact in a sub-additive manner. At first sight, this conclusion may appear to conflict with previous reports (Stein et al. 1989; Frassinetti et al. 2002), or indeed with some of our own findings (see Fig. 3a), which show

Fig. 2 Probability of bimodal stimulus detection as a function of stimulus level. Examples for bimodal visual/auditory, visual/tactile and auditory/tactile stimuli are shown. The raw, observed detection rates are shown in the *top row*, with fits for the additive and orthogonal model beneath. The stimulus intensities for the respective modalities are given by the x- and y-coordinates, respectively, and detection probabilities are given by the z-coordinate. The values for $x=0$ and $y=0$ correspond to false alarm rates, the values on the margins of the surface plots correspond to detection rates for unimodal stimuli, while the remaining points give detection rates at the various bimodal intensity combinations. For each model fit, the residuals (difference between observed and predicted detection rates) are shown to the right of the corresponding model fit



supra-additive interactions of detection probabilities at least for certain stimulus intensity ranges. However, this apparent conflict can be resolved provided that the relationship between the combined signal S_{xy} and detection probability p_{xy} exhibits an expansive non-linearity capable of “amplifying” the observed interaction. Signal detection theory assumes that the relationship between signal and detection probability is well approximated by a cumulative Gaussian function, which exhibits such an expansive non-linearity for small signal values. This expansive non-linearity makes it possible to formulate quite simple quantitative models which allow supra-additive enhancement of detection probabilities for certain ranges of stimulus intensities even if signals do interact sub-additively.

One such model, which turned out to be very successful indeed in accounting for the data, is the “Euclidean” or “orthogonal” model, which takes the form

$$p_{xy} = \Phi\left(\left((b_x \cdot \Delta x)^2 + (b_y \cdot \Delta y)^2\right)^{1/2} - \lambda\right) \quad (3)$$

Conceptually, the orthogonal model is a natural extension of Eq. 1 if we assume that the nervous system uses a “sensory metric space” to register and compare stimuli of

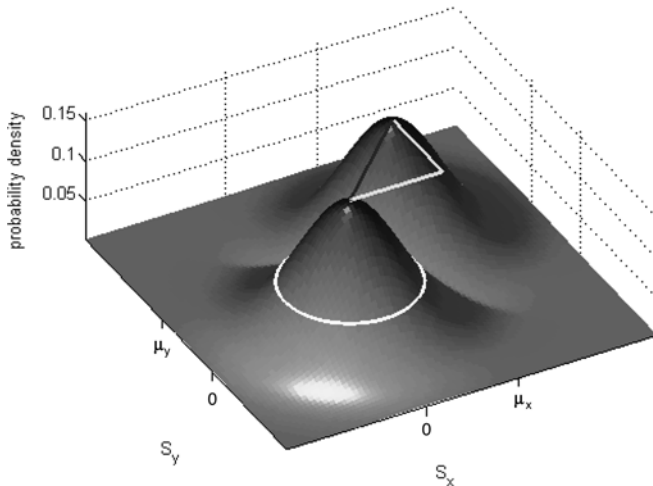


Fig. 3 The threshold signal detection model (cf Fig 1a) extended to multidimensional signals. The neural population signal representing a bimodal stimulus is now assumed to be a vectorial random variable $S=(s_x, s_y)$ with a bivariate Gaussian distribution. For simplicity, but without loss of generality, we again assume suitable normalization so that, in the absence of a stimulus (i.e. at background levels), the bivariate Gaussian has mean $(0,0)$ and unit variance. We assume that a stimulus in modality x shifts this distribution from its mean background level $\mu_{\theta}=0$ to a new mean $\mu_x=b_x \cdot \Delta x$ in the x direction, while a stimulus in modality y will shift the distribution from $\mu_{\theta}=0$ to $\mu_y=b_y \cdot \Delta y$ in the y direction. The distance the distribution is then shifted would then equal $\sqrt{(b_x \cdot \Delta x)^2 + (b_y \cdot \Delta y)^2}$ (length of the dark diagonal, Pythagoras’ theorem). In this scenario, detecting a stimulus amounts to deciding whether a particular signal value (s_x, s_y) is unlikely to have arisen from the “unshifted” (or “unstimulated”) distribution centered on $(0,0)$. The decision boundary for this task is effectively a circle of radius λ (shown as a white contour), and the decision criterion becomes whether $\sqrt{(s_x^2 + s_y^2)}$ exceeds λ or not

various types. The concept of a sensory metric space is quite abstract, and it is currently not clear how it might be implemented in the brain. Sensory metric spaces have nevertheless been used successfully to explain stimulus similarity judgments by human (Garner 1974) and non-human (Ronacher 1992) observers. In that context, the notion of a sensory space arises naturally as soon as animals acquire the ability to order stimuli according to intensity or size, so that the perceived size or intensity forms a “metric” which governs distances within the subjective sensory space. In the orthogonal sensory metric space envisaged here, different sensory modalities are thought to occupy separate and mutually orthogonal dimensions. Stimuli are represented as points in this space, and these points may move as the stimulus intensity in each modality/dimension changes. The dimensionality of this type of sensory space is potentially very large, as different submodalities (or sensory “channels”) within each sensory system may themselves occupy separate dimensions. The implications of such potentially high-dimensional sensory spaces for detection of cross-modal stimuli have so far been given mainly theoretical consideration (e.g. Drosler 2000) without recourse to experimental data.

Interpreting our results in such geometrical terms, Eq. 1 states that the probability of stimulus detection is directly related to the “distance” a stimulus level spans within the sensory space. For the bimodal case, Eq. 1 will extend to Eq. 2 if and only if the two sensory modalities “point in the same direction” (i.e. occupy the same dimension). The typically poor fit of the additive model indicates that this is not the case. Alternatively, two different sensory modalities may occupy separate, orthogonal dimensions. In that case, the neural signal S_{xy} representing a bimodal stimulus becomes a vectorial quantity. Its probability density function is then the bivariate normal distribution (see Fig. 3), and changes in the magnitude of this signal would have to be measured “along a diagonal” spanned by the components of the two-dimensional signal. The size of the signal would therefore be given by Pythagoras’ theorem, and Eq. 1 extends to Eq. 3 above.

The orthogonal model therefore has some theoretical appeal, but, much more importantly, unlike the additive model it also produced an excellent fit of the data. In not a single case was the residual deviance of the orthogonal model larger than expected by chance; in other words, the predictions of the orthogonal model are statistically indistinguishable from the observed data. Examples of orthogonal model fits are shown in the *third row* of Fig. 2.

The orthogonal model thus provides an appealing description of the multisensory stimulus detection processes examined here, but it is nevertheless conceivable that alternative models may exist which provide equally good fits of the data, yet make very different assumptions about the underlying decision processes. For example, one could in principle conceive of sensory spaces with topologies that require metrics other than the orthogonal Euclidean metric. One possibility considered by Garner (1974) and Ronacher (1992) is that sensory spaces may

exhibit a general “Minkowski metric”. Minkowski metrics take the form $(x^n + y^n)^{1/n}$. The additive model described above can be thought of as a special case of a Minkowski metric with exponent $n=1$, while the orthogonal, Euclidean model represents another special case with $n=2$. However, any exponent greater than one could in principle lead to the sub-additive behavior exhibited by our data. We investigated the possibility of exponents other than 2 by fitting our data with a general Minkowski model of the form

$$p_{xy} = \Phi\left(\left((b_x \cdot \Delta_x)^n + (b_y \cdot \Delta_y)^n\right)^{1/n} - \lambda\right), \quad (4)$$

where the exponent n becomes a free parameter of the model, to be optimized by a maximum likelihood fit. Since this model has more free parameters than the orthogonal model, lower deviances in the fits are to be expected (and were indeed found), but in no case were the deviances of the general Minkowski model significantly lower than those of the orthogonal model at the 5% level. Also, the maximum likelihood estimates for the optimal exponent n were always close to 2 (mean 1.82, std 0.27), suggesting that Euclidean distances (i.e. metrics with exponent 2) provide indeed an adequate and parsimonious description of the sensory metric space in which the detection of multisensory stimuli appears to take place.

Another assumption one might question is whether the brain’s sensory space is necessarily orthogonal, or whether its dimensions might be skewed. Again it is possible to address this question by probing the data. If we assume that dimensions might span any arbitrary angle θ , then the detection model would take the form

$$p_{xy} = \phi\left(\left(\left(b_x \Delta_x + b_y \Delta_y \sin \vartheta\right)^2 + \left(b_y \Delta_y \cos \vartheta\right)^2\right)^{1/2} - \lambda\right). \quad (5)$$

When we fitted this “free angle” model, we always obtained best fit angles close to 90° (mean 89.98° , std 8.5°) and in no case did introducing the free angle parameter lead to a significantly improved fit.

Finally, one could question whether the decision process does indeed occur after the two signals representing the individual modalities have been processed individually, and entertain the alternative hypothesis that detection decisions are made effectively separately and independently within each sensory modality. In that case, a subject would report the detection of a bimodal stimulus simply when she detected a stimulus either in modality 1 or in modality 2 (or both). Using boolean algebra and Eq. 1 it is easy to show that, under the assumptions of this “independent decision model”, the bimodal detection probability should obey the equation

$$p_{xy} = 1 - (1 - \Phi(b_x \cdot \Delta_x - \lambda)) \cdot (1 - \Phi(b_y \cdot \Delta_y - \lambda)). \quad (6)$$

This independent decision model produced fits that were superficially similar to those of the orthogonal model, but the deviances of the independent model fits were larger than those of the orthogonal model in 14 out of 17 cases, suggesting that the orthogonal model does fit the data overall significantly better ($p=0.0148$, Wilcoxon sign rank test).

In summary we conclude that, of all the models we have considered, the orthogonal model provides the best description of the multisensory stimulus detection process. It is theoretically well motivated, parsimonious, and in excellent agreement with the experimental data.

Materials and methods

Experiments were approved by the local Ethical Review Committee of the Experimental Psychology Department of the University of Oxford, and conform to the ethical standards laid down in the 1964 Convention of Helsinki. Six healthy, normal adult volunteers (two female, four male) served as subjects for this study. Subjects were seated in a sound attenuated, dimly lit room. Holding a small stimulator between the thumb, index and middle finger of their left hand, they rested the left hand, palm facing upward, on their left thigh, so that the stimulator was about 35 cm away from the subject’s face. The stimulator weighed 13.9 g and consisted of a vibro-tactile stimulator (Oticon BC-461), a small loudspeaker (Sony MDRE818LP) and an LED (RS components 247-1151), mounted together so that the three components were all aligned in one line of sight when the subject held the device as described. The vibrator, speaker and LED were each controlled through digital signal processing hardware (TDT system 3, Alachua, FL). Unlike other authors who investigated cross-modal effects on subjective brightness judgments (Stein et al. 1996; Odgaard et al. 2003), we did not carry out experiments in complete darkness or dark-adapt our subjects. Since we intended our subjects to detect changes in the sensory scene, it was sufficient to ensure that the room illumination was dim enough that the room light reflected off the body of the LED was a negligible fraction of the total amount of light emanating from the body of the LED.

In their right hands, the subjects held a response switch, which they were instructed to push to the left when they did detect the presence of a stimulus over and above the constant background (regardless of whether they considered the stimulus bimodal or unimodal) or to the right when they did not detect the stimulus. The next stimulus was presented automatically ca 0.7–1 s after the response to the previous stimulus had been registered. Data were collected in blocks of 128 trials, four blocks per session, and the subjects were allowed to take short breaks between

blocks to prevent fatigue and lapses of concentration. To allow the subjects to familiarize themselves with the task the first session was considered a training session. Data were collected from the following 2–3 sessions and were pooled to yield data sets comprising 1,024–1,536 responses.

Stimuli consisted of 100-ms intensity steps, presented against constant background intensities. Auditory stimuli consisted of broadband noise (0.1–12.5 kHz) with a background intensity of 51 dB SPL. The LED was held at a constant background luminance of either 6.8, 10.3 or 15.4 cd/m², and visual stimuli were delivered as 100-ms pulses of increased luminance. Vibro-tactile stimuli were 150-Hz sinusoidal vibrations with background RMS amplitudes between 16.2 and 48.5 N. The vibro-tactile device did not emit audible sounds. All stimuli were ramped on and off linearly with a 20-ms rise/fall time. For bimodal stimuli, both active modalities were always ramped on and off together in precise temporal synchrony. Stimulus intensities are reported here as fractional increments in signal amplitude above background. Stimulus levels were varied in eight equal steps between 0 and 0.14 (i.e. 0–14% above background) in the visual or auditory modalities. Given the somewhat lower sensitivity observed in the somatosensory modality, the vibro-tactile intensity steps covered a larger range, typically between 0 and 35% above background. Only one pair of modalities was tested in each session, i.e. each session contained 64 bimodal stimulus intensity combinations (including catch trials where stimulus intensity was zero in both modalities, and seven “unimodal” conditions where only one of the intensities was non-zero). Repeat presentations of these 64 combinations were randomly interleaved in each testing block. Background intensities were held constant in any one experimental session, but visual and vibro-tactile background intensities were occasionally varied from one session to the next to test whether results would generalize across background intensities. From the six subjects we collected a total of 17 different datasets (referred to as “cases” in the results section), covering a variety of modality combinations and various background intensities. Weber’s law would lead us to predict that the responses should be largely independent of background level and indeed we observed no significant changes in model parameters as a function of background level.

Model parameters were fitted to the data using maximum likelihood estimation. The number of reported stimulus detections is thought to be binomially distributed, with an underlying detection probability p_{detect} which is an unknown function of stimulus levels x, y . (Note that we do not distinguish between correct or mistaken detection reports.) The objective of the modeling process is to find a good approximation to this unknown function $p_{detect}(x, y)$. To determine the likelihood of a particular model given the data we calculated, for each stimulus level (x, y) , the probability p_O of observing O reported detections given that the model predicts an expected number of detections equal to $p_{success}(x, y)$ times the number of stimulus presentations N . The likelihood L is then simply the

product of the p_O over all levels (x, y) . Numerical optimization algorithms (Matlab optimization toolbox, Mathworks, Natick, MA) were used to determine the model parameters that maximize L .

The goodness of fit of the optimized model is then assessed using a deviance statistic (Berry et al. 2001). The deviance is defined as $D=2\cdot\log(L/L_0)$, where L is the likelihood of the model as defined above, and L_0 is the likelihood of the “saturated model” (i.e. a model that assumes that for each stimulus parameter combination the probability of stimulus detection equals the observed mean detection rate, i.e. $p_{detect}=O/N$). The deviance has two important statistical properties. Firstly, D is χ^2 distributed with a number of degrees of freedom that equals the number of stimulus parameter combinations tested (in our case 64) minus the number of free parameters in the model k . Hence, if $1-\chi^2(D, 64-k)<\alpha$, then we can reject the hypothesis that the model provides an adequate fit of the data at significance level α . Secondly, when comparing two competing models M_1, M_2 with deviances D_1, D_2 , and number of free parameters $k_1, k_2, k_2>k_1$, then the difference D_1-D_2 is also χ^2 distributed with degrees of freedom $df=k_2-k_1$. Hence, we can conclude that M_2 fits the data significantly better than M_1 if and only if $1-\chi^2(D_1-D_2, k_2 - k_1)<\alpha$.

Discussion

It is instructive to consider our results in the wider context of other previous studies. Let us first consider the failure of the additive model. In a series of recent papers, Ernst, Banks and colleagues (Ernst and Banks 2002; Hillis et al. 2002) studied the performance of human observers who were asked to combine visual and tactile sensory information in order to estimate the size of an object. They could account for their observations very well on the basis of a model that rests on the assumption that the brain forms a weighted sum of sensory signals across sensory modalities. So it appears that in the context of Ernst and Banks’ experiment, additive models apply, while in ours they seem to fail. This apparent discrepancy can be easily understood when one considers an essential difference in the tasks the subjects are asked to perform in each case. In Ernst and Banks’ experiment, the subjects are required to use two separate sensory modalities (vision and touch) to derive two independent estimates of a single physical property (the height of a raised ridge). The subjects will approach this task under the natural assumption that the perceived size of the object should be independent of whether the object is explored by sight or by touch. Consequently there is an expectation of a very tight correlation between the corresponding visual and tactile sensory signals in the observer’s mind. If this expectation is justified, then the averaging implied in the weighted sum model used by Ernst and Banks gives an optimal combined estimate of the single underlying quantity. The situation in the task we studied here is very different. In our task, it would be wrong and unjustified for the subjects

to assume a strong correlation between the intensity of the visual, auditory and tactile signals. Neither in the real world nor in our experiment does the fact that an object is loud automatically imply that it will also be bright. Whenever two sensory parameters are not necessarily highly correlated, each sensory modality must operate independently to signal a separate aspect of the stimulus. The signals from each sensory modality then cannot simply be combined additively into a single measure, but must be kept separate. However, if they are indeed processed separately, then one would not automatically expect that detection probabilities for bimodal stimuli would ever exceed the sum of the unimodal detection probabilities¹, which makes the original observations by Stein and colleagues (Stein et al. 1988) so intriguing.

The strength of the orthogonal model is that it allows the signal values for each modality to remain represented independently along separate dimensions. And while it combines the signals associated with each sensory modality in a sub-additive manner, it nevertheless permits supra-additive probability enhancement provided stimulus intensities remain within certain ranges. It is therefore not only entirely compatible with previous reports of supra-additive detection (Stein et al. 1988, 1989; Frassinetti et al. 2002), it can also be used to make precise predictions regarding the range of stimulus intensities for which supra-additive enhancement can occur. Figure 4a shows the difference between predicted bimodal stimulus detection probabilities (from Eq. 3) and the sum of the unimodal detection probabilities (from Eq. 1) using the model parameters estimated from the fit to the dataset shown in the middle column of Fig. 2. The white contour line indicates the stimulus intensities for which $p_{xy} = p_x + p_y$. Supra-additive facilitation occurs only when the stimulus parameters fall within the bounds of this white contour.

The orthogonal model also allows us to re-examine some of the previously formulated rules of multisensory integration in a more quantitative manner. For example, as mentioned in the introduction, observations from single unit recordings in the superior colliculus led Stein and Meredith (1993, p 143) to propose the “law of inverse effectiveness”, which states that “maximal enhancement occurs with minimally effective stimuli”. Stein (1988) and coworkers have also published experimental evidence suggesting that other rules of multisensory integration, namely the principles of temporal and spatial coincidence,

¹ Assume that the processing of the two sensory modalities occurred “separately” in the sense that these processes are statistically independent, as we have assumed in the formulation of the “independent model” (Eq. 6) above. Then, if the detection probabilities in each modality are p_x and p_y , the probability of *failing* to detect the stimulus would be $(1-p_x)$ or $(1-p_y)$, respectively. Failure to detect a bimodal stimulus occurs when the observer fails to detect the stimulus in modality X *and* in modality Y. Under statistical independence, the probability of failure in both modalities would equal the product of the individual failure probabilities, i.e. $(1-p_x)(1-p_y)$. The bimodal detection probability should then equal one minus this bimodal failure probability, i.e. one would expect that $p_{xy} = 1 - (1-p_x)(1-p_y) = p_x + p_y - p_x p_y$. Given that $p_x p_y > 0$, this must always be less than the sum of unimodal detection probabilities $p_x + p_y$.

would seem to apply equally to electrophysiological and psychophysical data, and one might assume that the law of inverse effectiveness should be similarly valid both for electrophysiology and behavior. Quantitative predictions from the orthogonal model allow us to examine this assumption explicitly. In their neurophysiological studies, Stein and colleagues (1993) measure enhancement simply as the percentage by which the bimodal response exceeded the larger of the two corresponding unimodal responses. By applying this enhancement index to our modeling results, which show that unimodal and bimodal stimulus detection probabilities are governed by Eqs. 1 and 5, respectively, we can predict the percent enhancement in detection probability for any particular combination of stimulus intensities. The resulting “behavioral enhancement function” is shown in Fig. 4b. Clearly, for increasingly large stimulus intensities the percentage enhancement rapidly decays to naught, as the law of inverse effectiveness predicts. However, our results clearly indicate that percent enhancement is a non-monotonic function of stimulus level, i.e. enhancement also declines substantially when stimulus effectiveness falls below detection rates of a few percent. At least for behavioral data it therefore seems that the law of inverse effectiveness holds only as long as stimuli do not become “too ineffective”. Maximal enhancements occur at stimulus intensities at which unimodal detection probabilities lie roughly between 10 and 30% (compare Figs. 1 and 4b).

This study provides a clear demonstration that the rules governing the psychophysics of multisensory integration can be formulated using precise, yet still fairly simple, mathematical models. So far we have only investigated the detection of multisensory stimuli that are presented in temporal synchrony and spatial register. A comprehensive mathematical reformulation of the principles of multisensory integration will require further experimental and analytical work, so that additional spatial and temporal variables can be introduced into future models. This development of quantitative rules for multisensory integration is likely to have useful technological applications. For example, formulae like those developed here should make it possible to adapt intensity levels of bimodal visual/auditory warning signals automatically to ongoing background noise and brightness levels. Thereby, warning signals could be adjusted so as to guarantee detection at some specified high probability, without making them so intense that they add unnecessarily to the auditory or visual noise load.

The scientific importance of our results, however, lies in the challenges the results raise for future research in the sensory neurosciences. Our orthogonal model arises naturally from the concept of a Euclidean sensory space in which different sensory modalities occupy separate dimensions, yet how the central nervous system implements computations in such an abstract stimulus space is still largely unclear. Although the statistical models described here account very nicely for the behavior of the organism as a whole, it is not immediately obvious how they relate to the behavior of individual neurons in

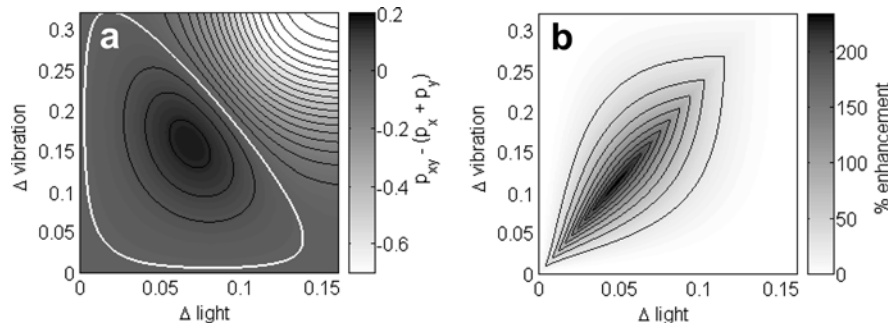


Fig. 4 **a** Difference between the bimodal stimulus detection probability p_{xy} (as given by the orthogonal model) and the sum of the two corresponding unimodal detection probabilities. The function shown is based on model parameters estimated for the data shown in the middle column of Fig. 2, and equivalent functions for other subjects or modality combinations will look very similar. Contour lines are shown at intervals of 0.05. The white contour line

shows $p_{xy} - p_x + p_y = 0$. The bimodal detection probability exceeds the sum of the unimodal detection probabilities if and only if stimulus levels (x, y) fall inside the white contour. **b** Multisensory enhancement (relative to the larger of the two unimodal detection probabilities) as a function of stimulus level, calculated from the detection probabilities given by the orthogonal model shown in the middle column of Fig. 2

multisensory sensory areas of the brain. For example, Meredith and Stein have documented individual multisensory neurons for which the largest superadditive gains are seen when unimodal stimuli are at, or even below, levels sufficient to evoke a response. King and Palmer (1985) have made similar observations in the superior colliculus of the guinea pig. These observations of enhancement at sub-threshold stimulus levels may appear to conflict with our finding that, psychophysically, enhancement declines when stimuli become too weak. However, these observations may easily be reconciled if we assume, for example, that the thresholds for these particular neurons are significantly higher than those seen behaviorally, or if these neurons make only a negligible contribution to the behavioral choice, or both. Stein and colleagues (Stein et al. 1988, 1989) have argued that there are strong similarities between electrophysiological and psychophysical results in studies of multisensory integration. However, a direct, quantitative comparison between electrophysiological and psychophysical results is only possible if the electrophysiological data are quantified using specific methods that facilitate such comparisons. Such methods have been employed very successfully in the study of the visual system. For example, by recording both psychophysical performance and neural responses in the same, behaving animal, and by quantifying neural responses in terms of “neurometric functions” (Tolhurst et al. 1983), Britten et al. (1992) were able to make a direct comparison of neural and behavioral thresholds for animals engaged in visual motion discrimination tasks. While in a number of cases, neurometric functions of neurons in MT closely resembled psychophysical performance, Britten and colleagues also observed neurons whose neurometric thresholds were either significantly above or below of the values obtained behaviorally.

Here we assume that the response properties of individual multisensory neurons may be quite diverse, and that each neuron will make only a small contribution to a neural population signal. Provided the population signal indeed reflects the summed or averaged activity of many neurons, the central limit theorem ensures that the

multivariate Gaussian distributions assumed in our analysis will arise naturally, regardless of the precise response properties of the individual neurons that constitute the population. In the light of the observation by Britten et al. (1992) that neurometric functions of some single neurons in MT can closely mirror psychophysical performance, one might question these assumptions and propose that the population code on which perceptual decisions are based could, in principle, be carried by only a very small number of neurons. However, in an elegant follow-up study, Britten and colleagues (1996) were able to estimate the size of the contribution made by individual neurons to the animal’s overall perceptual decision on a trial-by-trial basis. This contribution was typically so small as to be barely measurable, even for neurons whose sensitivity was well matched to the psychophysical data. Subsequent analyses of these results suggested that, for motion direction signals in area MT at least, the neural signal on which behavioral decisions are based is likely to be carried in pools of at least 100 cells (Shadlen et al. 1996). Naturally, it would be highly instructive if the methodology developed by Britten and colleagues could be incorporated into future studies of the neurophysiology of multisensory interactions.

In principle it is conceivable that the separate dimensions of the multidimensional neural population signal might be represented by largely separate but simultaneously active subpopulations of predominantly unimodal neurons. But while it is relatively easy to envisage biologically plausible representations of multivariate Gaussian signals in the brain, it is much harder to formulate testable hypotheses on how the brain subsequently performs computations on the magnitudes of these signals. Theoretical considerations discussed elsewhere (Green and Swets 1974; Bishop 1995; Wickens 2002) suggest that the orthogonal model described here may implement an optimal decision strategy for the combination of information across sensory channels, provided that the decision criterion λ is set appropriately to reflect the perceived relative costs and benefits that a subject attributes to correct or mistaken detection decisions. As

yet we have no clear indication on how the brain sets and implements such decision criteria, nor do we know where in the brain detection decisions for multisensory stimuli are ultimately computed. Multisensory neurons in the superior colliculus might be considered plausible candidates for this function, as might be any number of multisensory cortical areas. Advances in multielectrode recording techniques should make it possible in the near future to record from sufficiently large samples of neurons in animals trained in bimodal detection tasks to estimate population signals on a trial-by-trial basis, and to use these estimates to predict the animal's responses. Such experiments would bring us a lot closer to bridging the gap from neuron to behavior.

Acknowledgements We are grateful to Dr John Bithell for guidance on statistical methodology.

References

- Berry G, Matthews JNS, Armitage P (2001) Statistical methods in medical research. Blackwell Science Inc, Oxford
- Bishop CM (1995) Neural networks for pattern recognition. Clarendon Press, Oxford
- Britten KH, Shadlen MN, Newsome WT, Movshon JA (1992) The analysis of visual motion: a comparison of neuronal and psychophysical performance. *J Neurosci* 12:4745–4765
- Britten KH, Newsome WT, Shadlen MN, Celebrini S, Movshon JA (1996) A relationship between behavioral choice and the visual responses of neurons in macaque MT. *Vis Neurosci* 13:87–100
- Colonus H, Diederich A (2001) A maximum-likelihood approach to modeling multisensory enhancement. *NIPS*: 181–187
- Drosler J (2000) An n-dimensional Weber law and the corresponding Fechner law. *J Math Psychol* 44:330–335
- Ernst MO, Banks MS (2002) Humans integrate visual and haptic information in a statistically optimal fashion. *Nature* 415:429–433
- Frassinetti F, Bolognini N, Ladavas E (2002) Enhancement of visual perception by crossmodal visuo-auditory interaction. *Exp Brain Res* 147:332–343
- Garner WR (1974) The processing of information and structure. John Wiley & Sons, New York
- Georgopoulos AP, Schwartz AB, Kettner RE (1986) Neuronal population coding of movement direction. *Science* 233:1416–1419
- Green DM, Swets JA (1974) Signal detection theory and psychophysics. Krieger, New York
- Hillis JM, Ernst MO, Banks MS, Landy MS (2002) Combining sensory information: mandatory fusion within, but not between, senses. *Science* 298:1627–1630
- King AJ, Palmer AR (1985) Integration of visual and auditory information in bimodal neurones in the guinea-pig superior colliculus. *Exp Brain Res* 60:492–500
- King AJ, Schnupp JWH (1999) Multisensory convergence in neural function and development. In: Gazzaniga MS (ed) *The new cognitive neurosciences*. MIT Press, Cambridge, MA, pp 437–450
- Lovelace CT, Stein BE, Wallace MT (2003) An irrelevant light enhances auditory detection in humans: a psychophysical analysis of multisensory integration in stimulus detection. *Brain Res Cogn Brain Res* 17:447–453
- Meredith MA, Stein BE (1983) Interactions among converging sensory inputs in the superior colliculus. *Science* 221:389–391
- Odgaard EC, Arieh Y, Marks LE (2003) Cross-modal enhancement of perceived brightness: sensory interaction versus response bias. *Percept Psychophys* 65:123–132
- Ronacher B (1992) Pattern recognition in honeybees: multidimensional scaling reveals a city-block metric. *Vis Res* 32:1837–1843
- Shadlen MN, Britten KH, Newsome WT, Movshon JA (1996) A computational analysis of the relationship between neuronal and behavioral responses to visual motion. *J Neurosci* 16:1486–1510
- Stein BE, Meredith MA (1993) *The merging of the senses*. MIT Press, Cambridge
- Stein BE, Huneycutt WS, Meredith MA (1988) Neurons and behavior: the same rules of multisensory integration apply. *Brain Res* 448:355–358
- Stein BE, Meredith MA, Huneycutt WS, McDade L (1989) Behavioral indexes of multisensory integration. *J Cogn Neurosci* 1:12–24
- Stein BE, London N, Wilkinson LK, Price DD (1996) Enhancement of perceived visual intensity by auditory stimuli: a psychophysical analysis. *J Cogn Neurosci* 8:497–506
- Tolhurst DJ, Movshon JA, Dean AF (1983) The statistical reliability of signals in single neurons in cat and monkey visual cortex. *Vis Res* 23:775–785
- Wallace MT, Meredith MA, Stein BE (1992) Integration of multiple sensory modalities in cat cortex. *Exp Brain Res* 91:484–488
- Wickens TD (2002) *Elementary signal detection theory*. Oxford University Press, Oxford