Human trimodal perception follows optimal statistical inference

David R. Wozny

Biomedical Engineering IDP, UCLA, Los Angeles, CA, USA

Ulrik R. Beierholm

Computation and Neural Systems, California Institute of Technology, Pasadena, CA, USA

Ladan Shams

Department of Psychology, UCLA, Los Angeles, CA, USA

Our nervous system typically processes signals from multiple sensory modalities at any given moment and is therefore posed with two important problems: which of the signals are caused by a common event, and how to combine those signals. We investigated human perception in the presence of auditory, visual, and tactile stimulation in a numerosity judgment task. Observers were presented with stimuli in one, two, or three modalities simultaneously and were asked to report their percepts in each modality. The degree of congruency between the modalities varied across trials. For example, a single flash was paired in some trials with two beeps and two taps. Cross-modal illusions were observed in most conditions in which there was incongruence among the two or three stimuli, revealing robust interactions among the three modalities in all directions. The observers' bimodal and trimodal percepts were remarkably consistent with a Bayes-optimal strategy of combining the evidence in each modality with the prior probability of the events. These findings provide evidence that the combination of sensory information among three modalities follows optimal statistical inference for the entire spectrum of conditions.

Keywords: multisensory integration, cross modal, Bayesian inference, ideal observer, cross-modal illusion, trimodal perception, causal inference

Citation: Wozny, D. R., Beierholm, U. R., & Shams, L. (2008). Human trimodal perception follows optimal statistical inference. *Journal of Vision, 8*(3):24, 1–11, http://journalofvision.org/8/3/24/, doi:10.1167/8.3.24.

Introduction

When we walk on the street, we receive somatosensory stimulation from our feet, proprioceptive information about our body parts, visual information from various sources (other people, cars, buildings, trees, ground, sky, etc.), auditory stimulation from the various sounds caused by the cars, birds, our own footstep, olfactory stimulation from the smell of car exhausts, and so on. The vast majority of perception research has focused on studying one modality at a time. Here we investigate trisensory processing in human observers by simultaneously presenting visual, auditory, and tactile stimuli and by probing observers' perception simultaneously in multiple modalities. We examine how the three sensory signals interact, under which circumstances the signals from different modalities get bound together (integration), and when they are processed independently (segregation). In the processing of sensory information, the nervous system must infer what are the causal events that actually happened in the world based on the sensory information available and prior knowledge about the world. By studying how human observers perceive the stimuli in various stimulus conditions, we gain insight into the inference rules that govern perceptual processes. Given the variability in subject responses, and the stochastic nature of individual neurons, it is reasonable to believe that a statistical inference process is involved. We therefore compare human perception with a normative model based on Bayesian statistical inference. The Bayesian model performs inference not solely based on the information available but also relies on prior experiences. Multisensory information and prior experiences are combined in a way that, although not always veridical, is optimal in providing a lower bound for the error. Examining whether the Bayesian model makes the same errors (illusions) as human observers can be particularly informative. The mere fact that illusions occur underscores the need to investigate the underlying inference rules leading to the illusions. By examining conditions of conflicting sensory information, we gain insight about the integration–segregation dimension of multisensory processing.

The traditional model of sensory cue combination (Bülthoff & Mallot, 1988; Yuille & Bülthoff, 1996) has been successful in accounting for sensory integration, such as visual–haptic integration (Ernst & Banks, 2002; Hillis, Ernst, Banks, & Landy, 2002), visual–auditory integration (Alais & Burr, 2004), sensory-motor integration

doi: 10.1167/8.3.24 Received August 24, 2007; published March 27, 2008 ISSN 1534-7362

(Ghahramani, 1995; Ghahramani, Wolpert, & Jordan, 1997), and visual–proprioceptive integration (van Beers, Sittig, & Denier van der Gon, 1999). However, the traditional model assumes complete integration and therefore does not account for the vast range of situations in which the signals do not get integrated or are only partially integrated. It has recently been shown that Bayesian inference can account for a spectrum of integration and segregation in an auditory–visual tasks (Körding et al. 2007; Roach, Heron, & McGraw, 2006; Shams, Ma, & Beierholm, 2005) and a visual–haptic task (Bresciani, Dammeier, & Ernst, 2006). Here we will examine whether a normative model based on Bayesian inference can account for interactions among three modalities and explain the spectrum of auditory–visual–tactile sensory combinations. In our experiment, auditory, visual, and tactile stimuli were presented simultaneously to human observers. Zero, one, or two pulses for each modality were presented during each trial, and the observers were instructed to report the number of pulses that they perceived in each modality (providing three responses in each trial). In this paradigm, all three types of stimuli presented are task relevant. Alternatively, any of the modalities could be considered a distractor for the perception of each of the other modalities. Between two sensory modalities, the effect of distractor stimuli influencing the perception of a task-relevant stimulus has been shown extensively in the past. Shams, Kamitani, and Shimojo (2000, 2002) showed that multiple auditory beeps could cause a single flash to be perceived as two flashes. Bresciani et al. (2005) and Hötting and Röder (2004) demonstrated that multiple auditory tones influence the perceived number of tactile taps to the fingertip. Tactile distractor stimuli were shown to influence perceived number of flashes (Violentyev, Shimojo, & Shams, 2005). Sanabria, Soto-Faraco, and Spence (2005) investigated the effects of bimodal distractor stimuli on the perception of auditory apparent motion and showed that bimodal distractors had a greater influence on apparent motion detection than any of the unimodal distractors. In our study, two-way as well as three-way interactions among touch, vision, and hearing were explored in a variety of conditions ranging in the degree of discrepancy among the modalities. We then examined whether a Bayesian ideal observer that does not assume full integration can account for the combination of stimuli across all three modalities.

Methods

Participants

Twenty-four observers (13 female) with a mean age of 22 (range 17–31) participated in this experiment. One subject was left-handed. All of the subjects had normal or corrected-to-normal vision and did not have any auditory, tactile, or neurological disorders. Each participant signed a consent form approved by the UCLA IRB.

Stimuli

The visual stimulus was a uniform white disk 1.5 deg in diameter that was presented 7 deg below the fixation point. It was flashed for one frame (~10 ms) on an otherwise dark CRT computer monitor (refresh rate of 100 Hz) 0–2 times. Auditory beeps were played through computer speakers located on both sides of the computer monitor at a height equivalent to the visual stimulus. The auditory stimulus was a 10-ms ramped tone with 68 dB(A) sound pressure level and 3.5 kHz carrier frequency, also presented 0–2 times. Tactile stimulation was provided by a refreshable Braille cell consisting of a 2 × 4 array of plastic pins that extended 2 mm for 10 ms during stimulation and was also presented 0–2 times. A sample stimulus train for the case with one flash, two beeps, and one tap [$V = 1$, $A = 2$, $T = 1$] is shown in Figure 1.

The stimulus onset asynchrony was 60 ms for stimuli in each modality. The relative timing of the stimuli was adjusted such that the middle of the sequence of stimuli in each modality was aligned with that of the stimuli in the other modalities. A factorial design was used in which all combinations of 0–2 flashes, 0–2 beeps, and 0–2 taps (except for the 0 flash, 0 beep, 0 tap combination) were presented, leading to a total of 26 conditions. The experiment consisted of 10 trials per condition, totaling 260 trials presented in a pseudorandom order.

Participants sat 57 cm from the CRT monitor with their chins positioned on a chin rest. To mask the weak sound produced by the tactile device, the tactile device was placed inside a sound-attenuating chamber with a small sleeve for the subjects to place their left hand inside (palm down) and position their index finger on the Braille pins. In addition, continuous white noise was played through headphones at

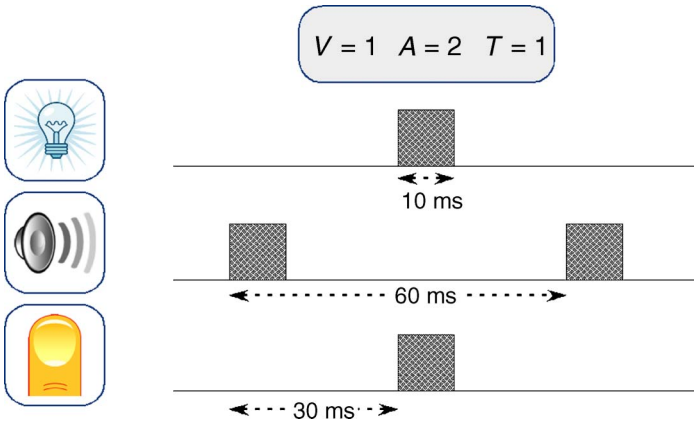

Figure 1. Timing of stimuli presentation for sample stimulus train condition [$V = 1$, $A = 2$, $T = 1$].

57 dB(A) to mask any residual noise produced by the Braille device. The participants were instructed to report the number of flashes, beeps, and taps perceived after each trial by entering three responses on a keypad with their right hand. Although the task may appear difficult at first, subjects were sequentially prompted on the monitor for the appropriate response, and since the order of responses did not change throughout an individual's session, they quickly became familiar with the task after a few practice sessions. Furthermore, subjects were also given an "escape" key they could press to back out and re-enter their responses prior to the start of the next trial. To further reduce the probability of noise introduced by memory, motor, or response bias, the order of responses was counterbalanced across subjects, and a large number of subjects (24) and total trials (6240) were acquired.

In order to familiarize participants with the task and to ensure that they are able to perform the task in unimodal conditions, they completed brief practice sessions prior to the experiment. Unimodal auditory, visual, or tactile trials were presented in random order, and after 30 trials feedback was provided about overall performance. A minimum of 90% correct performance for each modality was required in order to proceed with the experiment. Overall, most subjects required one or two practice sessions to achieve 90% accuracy in each modality. Two participants could not obtain the 90% correct after 4 practice sessions and were disqualified from participation.

Bayesian model

We tested whether the observed behavioral responses are consistent with a Bayesian observer model. A factor graph (Kschischang, Frey, & Loeliger, 2001) describing the statistical structure of the model is shown in Figure 2.

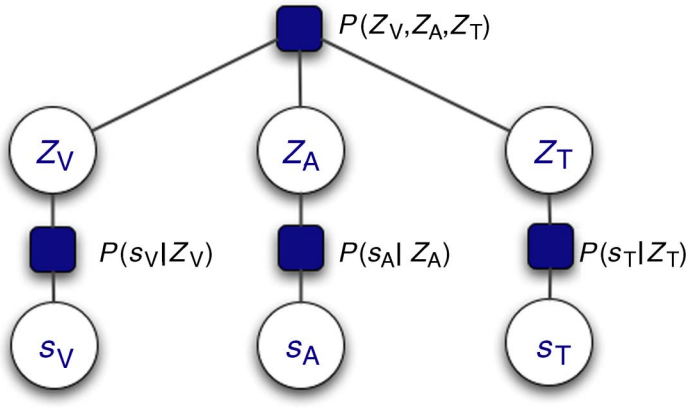

Figure 2. Factor graph for the Bayesian model, where circles represent random variables and squares represent probability functions over the random variables. Z_V, Z_A, and Z_T represent visual, auditory, and tactile sources, respectively; and s_V, s_A, and s_T represent visual, auditory, and tactile sensory signals.

Events (causes) in the environment lead to up to three types of stimuli, Z_V, Z_A, Z_T (denoting visual, auditory, and tactile stimuli, respectively). We assume that the sensory signals s_V, s_A, s_T are conditionally independent given the events Z_V, Z_A, Z_T. This common assumption is based on the fact that the signals of the different modalities are processed in separate pathways (up to the point where interactions occur) and are therefore corrupted by independent noise processes. A Bayesian observer would make the best possible guess about the causes given the sensory signals and prior probability of the causes. The best possible guess is achieved by optimal inference about posterior probabilities $P(Z_V, Z_A, Z_T \mid s_V, s_A, s_T)$ using Bayes' rule, which given the assumptions of the model described above simplifies to

$$P(Z_V, Z_A, Z_T|s_V, s_A, s_T) = \frac{P(s_V|Z_V)P(s_A|Z_A)P(s_T|Z_T)P(Z_V, Z_A, Z_T)}{P(s_V, s_A, s_T)}. \quad (1)$$

The likelihoods $P(s_V \mid Z_V)$, $P(s_A \mid Z_A)$, and $P(s_T \mid Z_T)$ were modeled as a normally distributed sensory signal centered about the true stimulus and corrupted by independent unbiased Gaussian noise with standard deviations σ_V, σ_A, σ_T, respectively. The joint prior probability $P(Z_V, Z_A, Z_T)$ was modeled as a multivariate normal distribution centered about its vector mean μ_{prior} with covariance matrix Σ_{prior}:

$$\mu_{prior} = \begin{bmatrix} \mu_V \\ \mu_A \\ \mu_T \end{bmatrix} \qquad \sum_{prior} = \begin{bmatrix} \sigma^2_{prior_V} & cov_{VA} & cov_{VT} \\ cov_{VA} & \sigma^2_{prior_A} & cov_{AT} \\ cov_{VT} & cov_{AT} & \sigma^2_{prior_T} \end{bmatrix}. \quad (2)$$

Note that the likelihood of each sensory signal given the cause is independent of each other, and integration of sensory information stems from the covariance in the prior distribution. The shape of the prior distribution determines the extent of integration between the sensory signals. A uniform prior (zero-covariance) would process each of the signals independently. A prior distribution with high density along the triagonal (strong covariance) assumes full integration of the signals. Our Bayesian inference model does not make either of these assumptions, and the variance and covariance can take on any positive values, allowing for full-integration, partial integration, as well as segregation.

We assume that for the given task, the prior knowledge does not differentiate among the modalities in terms of their mean numerosity, reliability, or interdependence, resulting in equal mean, variance, and covariance across modalities. This assumption was validated by fitting separate prior

parameters for each modality and obtaining similar parameter values across modalities (see Appendix A). Gaussians were used to model likelihood and prior distributions for their simplicity and mathematical convenience (Bresciani et al., 2006). We estimated the likelihood variances strictly from the observed unimodal conditions. Thus, the only parameters that were fitted to the data were parameters of the prior, resulting in only three free parameters: μ_{prior}, σ_{prior}, cov_{prior}.

In order to compare the human performance with that of the Bayesian model, we performed a Monte Carlo simulation of the generation of the causes and the inference about the stimuli (i.e., percepts). This procedure replicates the trial-to-trial variability observed in human responses. We simulate each stimulus condition 10,000 times and thus obtain a distribution of responses for each condition that can then be compared with the distribution obtained from human observers. Each simulated trial generates stochastic estimates for each sensory signal by sampling from the respective likelihood distribution. More specifically, the likelihood estimates are sampled from Gaussian distributions centered about the true signal corrupted by noise of width σ_A, σ_V, and σ_T for auditory, visual, and tactile modalities, respectively. The posterior is calculated according to Equation 1, using a prior distribution characterized by the three free parameters, μ_{prior}, σ_{prior}, and cov_{prior}. The appropriate $N \times 1$ prior mean vector and $N \times N$ prior covariance matrix is used in calculating the posterior distribution, where N is the number of stimulus modalities presented for that given simulated trial. This makes the assumption that subjects do not report any pulsations for an absent stimulus, i.e., no hallucinations or synesthesia. This assumption was indeed confirmed by the data. While we allowed reporting a non-zero number for an absent stimulus, subjects did so only in 2% of all possible unimodal or bimodal experimental trials, likely due to motor or memory errors. We assume that the observer tries to minimize the mean squared error (i.e., least squares loss function), and thus the optimal response would be the mean of the posterior distribution. Note that as we use normal distributions for the likelihood and priors, the posterior is also normally distributed, and taking the mean is equivalent to finding the maximum of the distribution. To produce a response based on this optimal estimate, for each modality, we then choose the response category (0, 1, or 2) nearest to the optimal estimate. These simulations result in a response distribution for each of the stimulus conditions. An optimization search is used to find the parameters of the prior distribution that minimize the mean squared error between the simulated responses and responses of human observers.

We compare performance of the Bayesian inference model to several alternative models. We report the coefficient of determination, $R^2 = 1 - SS_E/SS_T$, as the measure of goodness of fit between the model and the data, as well as the Bayesian Information Criteria (Burnham & Anderson, 2002).

Results

Cross-modal interactions

The response distributions of human observers in each of the 26 stimulus conditions are shown in Figure 3 with data points connected by broken lines. To facilitate the interpretation of the data, instead of showing the three-dimensional response matrices, the three one-dimensional marginalized distributions are shown for each condition. As can be seen, the responses are largely veridical in the unimodal and congruent bimodal and trimodal conditions.

However, as previously mentioned, it is the incongruent conditions that probe the statistical inference rules that the nervous system must use in resolving potential conflicts in sensory cues. In incongruent conditions, there are often considerable deviations from the veridical response in one or more of the modalities. In other words, the information in the task-irrelevant one or two modalities affects the response distribution in the task-relevant modality. However, not all differences between two response distributions necessarily reflect meaningful interactions between modalities, as they may be due to sampling error. For example, the change in the visual response between unimodal condition [$V = 1, A = 0, T = 0$] and bimodal condition [$V = 1, A = 2, T = 0$] may be either statistically insignificant or could correspond to a statistically significant modulation of visual perception by sound. Therefore, to find which of the changes between two conditions correspond to a statistically significant perceptual interaction, we performed the following analysis. We calculated d' discriminability index (Smith, 1982) for each of the modalities in each of the stimulus conditions and examined whether the change in sensitivity (d') between two conditions is statistically significant. Figure 4 shows the corrected t-statistics for each unimodal condition given in the graph title compared with all conditions that differ in the other two modalities. Test statistics were calculated by comparing the d' across subjects for perception in each modality between the unisensory condition (e.g., visual percept in [$V = 1, A = 0, T = 0$]) and corresponding bi-sensory (e.g., visual percept in [$V = 1, A = 2, T = 0$]) or tri-sensory condition (e.g., visual percept in [$V = 1, A = 2, T = 2$]) using a two-tailed paired t-test ($\alpha = 0.05$, Bonferroni corrected for 48 tests). The p-values are provided in gray scale with darker squares corresponding to lower p-values. Statistically significant tests are highlighted in red squares, and all were found to be in the right tail, shifting away from the veridical percept for that given modality (positive t-statistics). The first row of Figure 4 provides a statistical examination of illusory fission effects, in which the percept of a single pulse in one modality (e.g., a single flash or beep or tap) is changed into two pulses (two flashes, or two beeps, or two taps) when paired with two pulsations in one or both of the other modalities. The second row provides a statistical examination of illusory

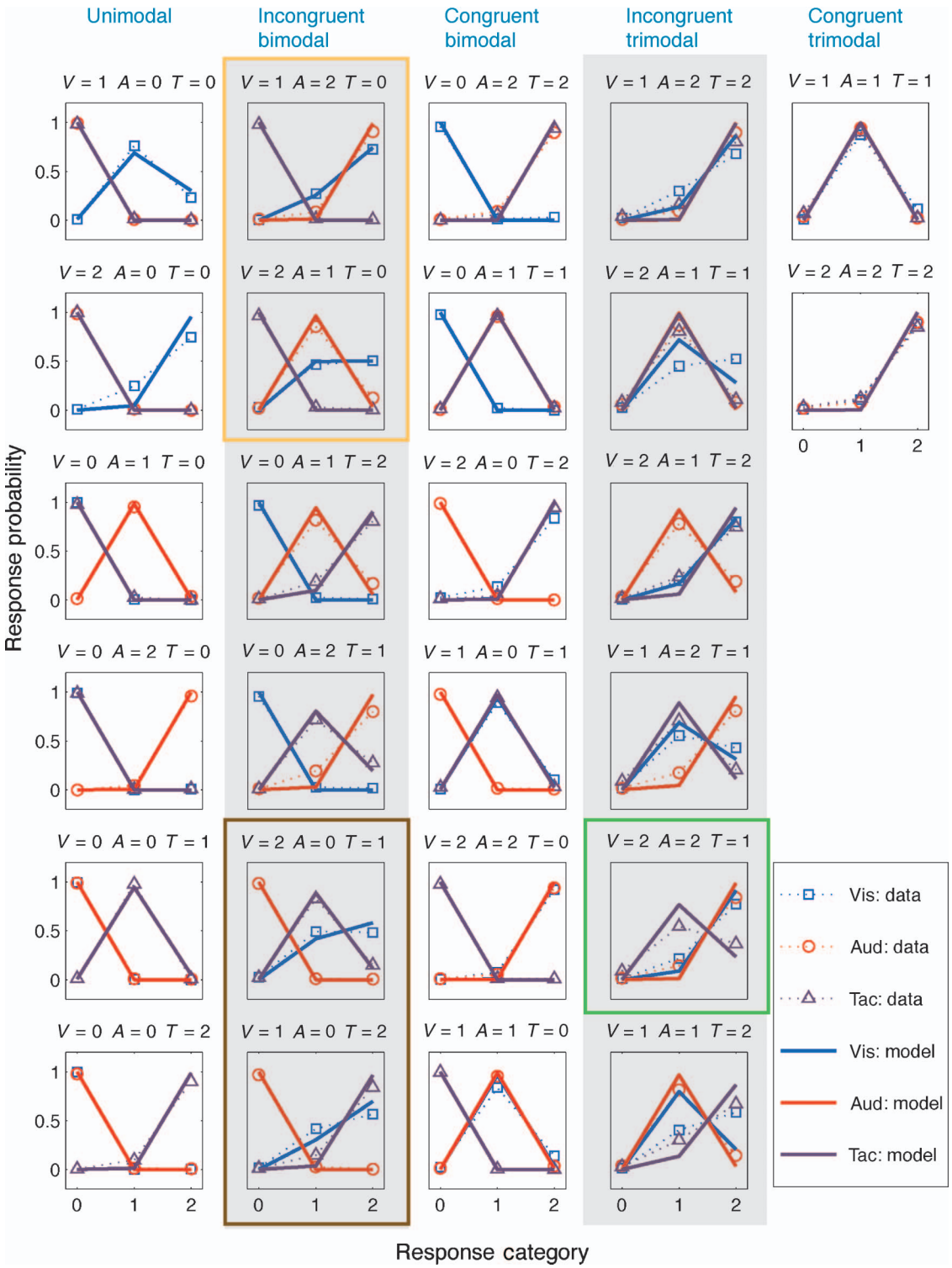

Figure 3. Marginalized response distributions fitted by the model in Figure 2, compared to behavioral data from 24 subjects for each of the 26 experimental conditions. Each graph title displays the number of visual (*V*), auditory (*A*), and tactile (*T*) stimuli presented for the condition. Horizontal axis shows the response category, and vertical axis shows the probability of response. Solid lines show the model predictions, and symbols connected with dashed lines with data points show the human observers' response distributions (see legend for shape and color details). Data are organized into columns describing the types of modalities presented. The orange box highlights sound-induced flash illusion, the brown box highlights touch-induced flash illusion, and the green box highlights one example of three-way interaction described in text. $R^2 = 0.95$.

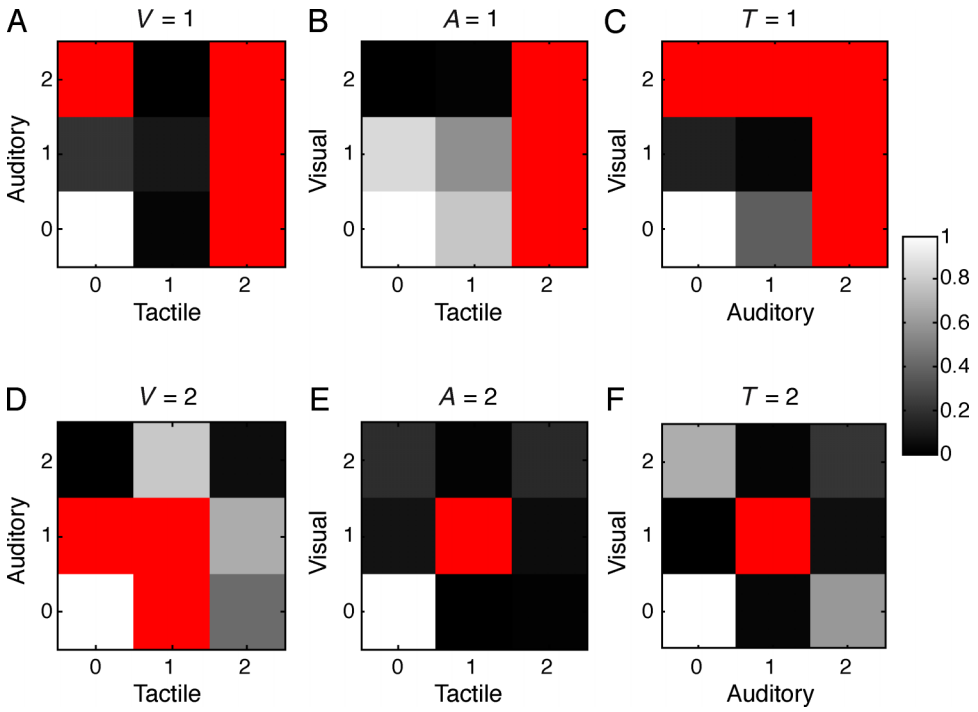

Figure 4. The *p*-values comparing *d′* measurements from 24 subjects. Each unimodal condition was compared to all alternative conditions using a paired *t*-test (*df* = 23). Significant results ($p < 0.05$, Bonferroni corrected) are shown as red squares. For each graph, the (0, 0) condition would be identical to the unimodal condition, so *p*-values were not calculated.

fusion effects, in which the percept of two pulses in one modality (2 flashes, or 2 beeps, or 2 taps) is changed into one when paired with one pulse in one or two of the other modalities.

By inspection of Figure 4, one can see there are many statistically significant interactions across the range of conditions. We list some examples to convey the complex dynamics that occur through the interaction of three sensory modalities, especially in the case of conflicting information. As pointed out below, a few of the interactions are already reported in the literature, but we were able to reproduce these as well as many more two- and three-way interactions all within a single experiment. By understanding the wide spectrum of interactions that occur, we can further appreciate if a single inference rule can explain the observed behavior.

Bimodal conditions

We first focus on the incongruent bimodal conditions. Figures 4A and 4D show that both tactile and auditory stimuli cause significant fission and fusion illusions in vision as have been previously reported (Shams et al., 2000, 2002; Violentyev et al., 2005). More specifically, in auditory–visual condition [$V = 1, A = 2, T = 0$], the sound-induced flash illusion (Shams et al., 2000) can be seen in the visual responses in which subjects report two flashes in a large fraction of trials (Figure 3 top plot in orange box) due to introduction of two beeps ($p < 0.001$; Figure 4A). A visual fusion illusion is found in the [$V = 2, A = 1, T = 0$] condition in which subjects report seeing one flash in a large fraction of trials as a result of pairing with 1 beep (Figure 3 bottom plot in orange box, and Figure 4D). Similar fission ($p < 0.001$) and fusion ($p < 0.001$) touch-induced visual illusions (Violentyev et al., 2005) are found in the [$V = 1, T = 2, A = 0$] and [$V = 2, T = 1, A = 0$] conditions (Figure 3 brown box, and Figures 4A and 4D). In addition to these previously reported illusions, we find weaker but statistically significant auditory and tactile illusions in some of the bimodal conditions. For example, in condition [$V = 0, A = 1, T = 2$], a touch-induced double-beep illusion occurs ($p < 0.001$, Figure 4B). Sound-induced and visually induced fission touch illusions also occur in conditions [$V = 2, A = 0, T = 1$] ($p < 0.001$, Figure 4C) and [$V = 0, A = 2, T = 1$] ($p < 0.001$, Figure 4C), respectively.

In congruent bimodal conditions, the accuracy is generally increased compared to unimodal conditions; however, these effects are not as large in magnitude and do not reach statistical significance (given our strict statistical criterion which corrects for multiple comparisons).

Trimodal conditions

In incongruent trimodal conditions, one modality is in conflict with the other two modalities. Whereas the fusion illusory effects in bimodal conditions are typically weak and do not reach statistical significance (except for the visual fusion, Figure 4D), the fusion effects are generally

stronger in the trimodal conditions and are statistically significant (Figure 4, bottom row). For example, a double beep paired with a single flash does not lead to a statistically significant single-beep illusion, but it does when it is paired with a single flash as well as a single tap (Figure 4E). Similarly, a statistically significant fusion of a double tap into a single tap only occurs in the presence of both a single flash and single beep (Figure 4F).

Other three-way interactions are observed when comparing the trimodal incongruent conditions with their respective bimodal incongruent conditions. These results are shown in Figures 5A–5F, with the bimodal condition appearing in the title, the modulating third modality is shown along the horizontal axis, and the squares represent the p-value associated with the change in d' from the original bimodal responses while showing the original bimodal pair along the vertical axis. Again, a paired two-tailed t-test was performed for each comparison, and p-values are provided in gray scale with darker squares corresponding to lower p-values. It is interesting to note that some of the significant effects are a result of the third modality introducing an illusory effect, similar to those in Figure 4. These are shown by red squares. The statistically significant changes that fall within the left tail (negative t-statistics) are shown in blue. The blue squares represent changes where the addition of a third modality resulted in a *decrease* in the initial illusion, i.e., a shift toward the veridical percept for that given modality. For example, comparing visual and tactile responses in $[V = 2, T = 1, A = 0]$ (Figure 3, top plot in brown box) vs. $[V = 2, T = 1, A = 2]$ (Figure 3, green box), we find that the modulatory effect of single tap on vision (leading to the visual fusion effect) seen in the former condition is significantly reduced in the later due to the introduction of the double beeps while increasing the rate of tactile fusion effects. The red square in the second column of Figure 5F indicates that adding two beeps to the $[V = 2, T = 1, A = 0]$ condition (shown in the title) results in significant increase in illusory tactile percepts (first row), and the blue square shows significant decrease in visual illusion (second row).

Comparison of trimodal conditions with their respective *congruent* bimodal conditions (see Figures 5G–5L) shows that the double pulses in the third modality consistently lead to fission effects in one or both of the congruent modalities (Figures 5G–5I), whereas the addition of a third single-pulse event to congruent bimodal events only leads to a tactile fusion effect (Figure 5L).

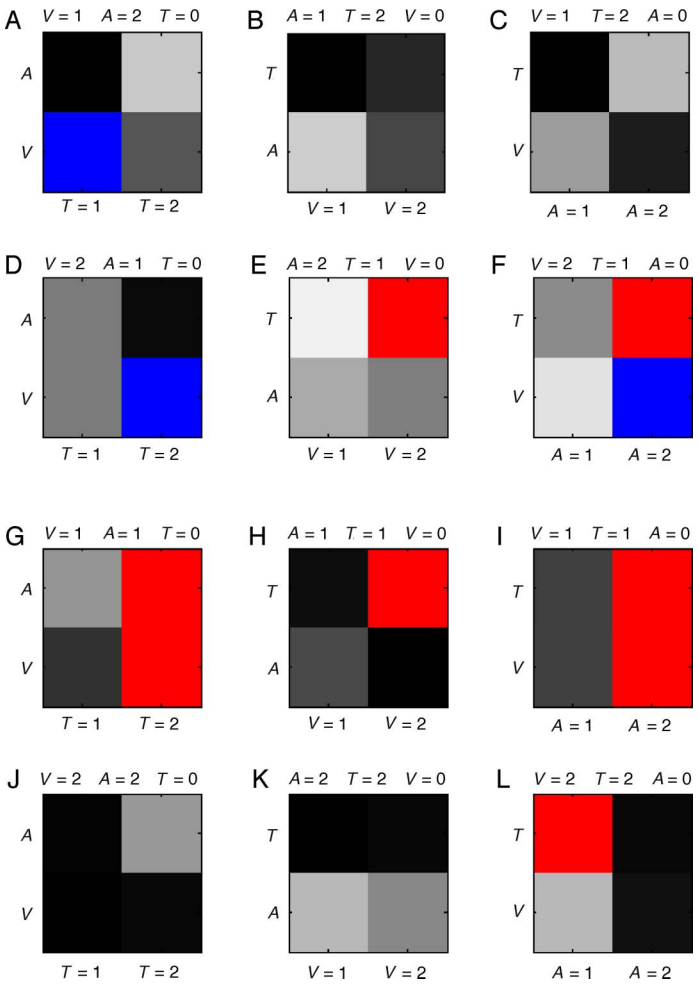

Figure 5. The p-values comparing d' measurements from 24 subjects between two conditions. Each bimodal condition was compared to all related trimodal conditions using a paired two-tailed t-test (df = 23). Significant results ($p < 0.05$, Bonferroni corrected) are shown as red squares for positive t-statistics, and blue squares for negative t-statistics. Each panel examines the change in d' in each of two modalities as a result of introduction of a third modality. Each row corresponds to the one of the two modalities, and each column corresponds to the number of pulses in the third modality. Each figure title displays the visual (V), auditory (A), and tactile (T), stimulation parameters for the bimodal condition (baseline comparison). For example, in panel g, there is not a significant change in the subjects' d' for either visual or auditory responses when adding a single tap, as shown by the gray scale boxes for both the auditory and the visual percepts in the column T = 1. In contrast, both auditory and visual d' significantly change when two taps are added.

Cross-modal interactions summary

In summary, Figure 4A shows that both auditory and tactile stimuli influence the number of flashes perceived. It appears that when the tactile and visual stimuli are congruent, the incongruent auditory information does not have as significant an effect. Figure 4B shows that the tactile but not the visual stimuli have a significant effect on auditory fission. In addition, Figure 4C shows that both vision and audition can produce tactile fission effects. Figure 4D shows that both tactile and auditory stimuli produce significant fusion illusory effects on visual stimuli either in the bimodal or trimodal conditions. In Figures 4E and 4F, significant fusion effects are only reported during the trimodal conditions when the two other stimuli are

Model	R^2 (incongruent)	BIC
Bayesian inference	0.8884	−455
Cross-validation	0.8746	−443
Independent	0.6792	−355
Forced fusion	0.5791	−326
Veridical	0.4340	−294

Table 1. Two measures of goodness-of-fit (R^2 and BIC) for the incongruent sensory conditions (8 × 12 = 96 data points; cf. Figure 3, columns 2 and 4).

congruent with each other. Thus, it appears that an incorrect decision can be made even for salient signals, like audition and touch, when enough evidence is constructed against the truth. Figures 5A–5F show that adding a third modality to an incongruent bimodal pair shifts responses toward the newly formed congruent pair. This shift can negate previous bimodal illusions and/or form new trimodal illusions. Figures 5G–5L shows that fission effects are stronger than fusion effects also in trimodal conditions. We will next compare the behavioral results with the Bayesian inference model to see how well it can account for such a vast array of interactions.

Comparison with Bayesian model

We compared the data obtained from human observers with the Bayesian model. The solid lines in Figure 3 represent the model fits. As can be seen, there is overall consistency between the model and the data. The model can explain more than 95% of the variance in the data ($R^2 = 0.95$). However, this measure of goodness-of-fit may be somewhat inflated since over half of the conditions are either unimodal or congruent stimulus presentations where the responses are close to veridical. As mentioned above, most interactions occur in incongruent conditions, and thus we are particularly interested in examining how the model would account for this data. The R^2 value for these conditions is presented in Table 1, along with the comparison values for three alternative models: independence, forced fusion, and veridical.

The independence model assumes that there is no interaction between the two modalities, and hence a flat prior (with zero covariance). The forced fusion model is the traditional model of cue combination (Alais & Burr, 2004; Ernst & Banks, 2002; Jacobs, 1999; van Beers et al., 1999) in which the two cues are assumed a priori to have been caused by a single event and always get completely fused. This model corresponds to a prior that is zero everywhere except for the triagonal (strong covariance), strictly enforcing the integration of sensory estimates. To assess the degree of interaction among the three modalities, we also tested a simple model, "the veridical model," in which subjects never make mistakes, and always report the correct number of pulsations for each modality. While this last model is not plausible, it provides a good estimate for the percentage of errors in subjects' responses during these illusory conditions. Since all three of the alternative models do not contain any free parameters, we also report in Table 1 the Bayesian information criterion (BIC) for each model. BIC discourages overfitting by incurring a penalty as an increasing function of the number of free parameters and penalizes free parameters more strongly than Akaike information criteria. In general, the model with the lower BIC score is preferred. To ensure that the parameters are not overfitting the data, we also fit the parameters to half of the subject data and tested the goodness of fit to the other half of the data. The R^2 of this cross-validation averaged over the two partings of the subjects is also shown in Table 1. The calculated unimodal variances and the optimized parameter values for the prior distribution are shown in Table 2 for group and individual subject fits. The relationship between the parameter values and the observed behavior will be explored in the Discussion section. The comparison of the model with individual subject's data also resulted in a good fit ($R^2 = 0.85 \pm 0.015$). This goodness of fit is not as high as that of the group data; however, this is to be expected due to the relatively small number of trials per condition for each subject.

In summary, the model accounts for the data well across all conditions, including the conditions where there is incongruence among the sensory modalities (cf. columns 2 and 4 of Figure 3), in which strong and varied two-way and three-way interactions occur. Notably, using only 3 free parameters, the model can explain 95% of the variance in subjects' responses. The Bayesian model that employs a trimodal interaction prior outperforms alternative models without an interaction prior. These results suggest that multisensory perceptual processing among two or more modalities can be well understood using a framework of Bayesian inference about sensory signals and their causes.

	σ_V	σ_A	σ_T	μ_{prior}	var_{prior}	cov_{prior}
Group	0.45	0.21	0.25	1.93	0.25	0.21
Individual	0.32 ± 0.031	0.13 ± 0.031	0.11 ± 0.030	1.38 ± 0.146	0.26 ± 0.012	0.17 ± 0.014

Table 2. Parameter values. The likelihood standard deviations for vision, audition, and tactile (σ_V, σ_A, and σ_T, respectively) were obtained from unimodal conditions. The three parameters of the prior were fitted to the data. First row shows parameter values for the data pooled across subjects. Second row shows the parameters' means and standard errors for individual subject fits ($N = 24$).

Discussion

Cross-modal interactions

We have investigated bi- and tri-modal perception by presenting stimuli in one, two, or three modalities and probing the perception in all three modalities at the same time. Unique to our experiment, all three of our modalities are considered focal signals, which allows us to simultaneously investigate the degree of interaction between visual, auditory, and tactile modalities. In addition to previously reported cross-modally induced visual illusions, we found cross-modally induced auditory and tactile illusions. In all modalities, the fission illusions (in which one pulsation is perceived as two) were generally stronger than fusion illusions (whereby two pulsations are perceived as one). Importantly, we also found three-way interactions in which the interaction between two modalities was modulated by the signal in the third modality. Or, equivalently, the percept in one modality was determined by the signals from both other modalities. These results clearly show that the interactions among modalities are ubiquitous and can occur in various combinations of modalities, and in various directions, often affecting the percept in both or all three modalities. The trials in which the percept in two modalities differed are examples of segregation. These trials constitute the majority of trials. Trials in which the percept in two (or more) modalities is the same likely reflect integration. Both fission and fusion illusions, in which the percept between two modalities is rendered equal (one pulsation in case of fusion, and two pulsations in case of fission), therefore represent integration. Previously, we have referred to trials in which the percept in two modalities are different but shifted toward each other (e.g., three beeps and one flash are perceived as three beeps and two flashes) as "partial integration." Because the discrepancy between the present signals in this experiment was small (and never exceeded one pulsation), we could not expect partial integration, however, partial integration has been shown to occur in the same and other tasks previously (Körding et al., 2007; Shams et al., 2005). Therefore, the data in this experiment span segregation and integration.

Comparison with Bayesian model

Most importantly, we found that a simple inference rule derived from Bayes' theorem can account for the entire set of data remarkably well. The values of the estimated parameters are quite reasonable for the task at hand. For example, the visual sensory standard deviation was estimated to be approximately twice as large as the tactile and auditory estimates. This is consistent with the general notion that vision has less temporal precision than the auditory and tactile systems. Participants' debriefings after the experiment also confirmed that they were most uncertain in their visual percepts, followed by tactile and auditory percepts, respectively. Inspecting the prior parameters, we find the variance to be fairly large indicating only a weak bias for any specific number of pulsations, and the covariance to be fairly large showing a strong inclination for integration (of signals that are close in numerosity) that gradually weakens as the discrepancy between the signals increase. This bias for integration increases with the number of pulsations due to the fact that the mean of the prior is well above zero. It seems reasonable to believe that life experiences would lead to building such a bias since it is more likely that two or more pulses simultaneously occurring across two or more modalities are due to related events rather than unrelated events. This might also explain why fission illusions are stronger than fusion illusions, where the greater number of sequences has a stronger influence for integration.

To date, few models have been able to account for the trial-to-trial variability in the psychophysical data (Körding et al., 2007; Stocker & Simoncelli, 2006). Similar to Körding et al. (2007) and Stocker and Simoncelli (2006), we are able to account for the trial-to-trial variability using a stochastic likelihood function. Similar to these models, we are also able to fit the priors using psychophysical data. As in Körding et al., we parameterize the prior distribution and fit the parameters by minimizing the difference between the subject responses and the responses predicted by the model. In contrast to Stocker and Simoncelli, who use a non-parametric distribution for the priors, we make a simplifying assumption of Gaussian distributions for the priors, resulting in a less flexible form for priors, however much fewer parameters, and thus somewhat more parsimonious. In doing so, we make the assumption that the underlying distributions are continuous functions up to the final read-out stage of the estimate, where a discrete decision is made. In everyday life, we often make categorical judgments such as which expressway lane is moving faster or what dinner to order off of the menu, and it is likely that continuous evidence processing is carried out up to the final discrete decision. Given that sensation is a continuous system, there has to be some kind of conversion from continuous to discrete somewhere along the path from sensation to decision making. Little evidence is available on how discrete variables are represented in the human brain. Electrophysiological studies have reported cells that show tuning curves specific for numerosity in the association cortex of anesthetized felines (Thompson, Mayers, Robertson, & Patterson, 1970) and the lateral PFC of awake behaving monkeys (Nieder, Freedman, & Miller, 2002). The neurophysiological data taken together with behavioral data and neural network models support analogue magnitude models for numerical representations (Dehaene & Changeux, 1993; Meck & Church, 1983; Moyer & Landauer, 1967; Nieder et al., 2002; Nieder & Miller, 2004; Restle, 1970; van Oeffelen & Vos, 1982; for

reviews, see Dehaene, 1992; Dehaene, Dehaene-Lambertz, & Cohen, 1998; Gallistel & Gelman, 2000). Therefore, it is possible that the nervous system utilizes magnitude representations of numerical information until a decision is made. Given the good fit of our model to the data, a Gaussian distribution appears to approximate the underlying distributions rather well.

In the real world, we are often surrounded by multiple objects and multiple sources of sensory stimulation. This leads to multiple temporally coincident sensory signals within as well as across modalities. While spatial congruency is a useful cue for binding, it is often not sufficient as the spatial resolution of most modalities is fairly poor. Thus, the decision on whether two or more signals correspond to the same object or different objects needs to take the structural consistency of the stimuli into account. The findings of this study illustrate this phenomenon well. Although the relative timing and spatial position of the stimuli is fixed across trials, we observe a range of sensory combination from integration to segregation as a function of structural discrepancy among the stimuli. The Bayesian inference model accounts for this spectrum of integration by allowing dissimilar stimuli to create shifts in perception of each other, yet still allowing them to be separate estimates. The higher the variance in the sensory estimate of the stimuli (i.e., likelihood), the greater the influence the prior will have in inferring the event.

Conclusions

We introduced a trimodal experiment that simultaneously presented subjects with visual, auditory, and tactile pulsations and probed their numerosity perceptions in all three modalities. This allowed us to examine multiple sensory interactions within a single experiment, replicating some previously reported cross-modal illusions, as well as reporting a wide spectrum of novel two- and three-way statistically significant interactions. A single normative model based on Bayesian inference was shown to account for the entire range of phenomena observed in the behavioral data. In contrast to the traditional models of cue combination, this model does not make an a priori assumption of integration and allows independent causes for the observed cues.

Appendix A

The goodness of fit obtained for the 9-parameter model is slightly better than the 3-parameter model (R^2 = 0.97 vs. R^2 = 0.95). Importantly, the similar parameter values across the modalities in the 9-parameter model (Table A1) confirms our assumption of equal values across modalities for the mean, variance, and covariance, resulting in 3 free parameters (see Methods section).

	μ_{prior}	var_{prior}	cov_{prior}
Vision	1.92	0.29	0.19 (VA)
Audition	1.85	0.22	0.22 (AT)
Tactile	1.85	0.24	0.21 (TV)

Table A1. Optimized parameter values for the 3D prior allowing for separate mean, variance, and covariance values for each modality.

Acknowledgments

We thank Konrad Kording for his valuable contribution to the development of the model, the insightful discussions, and the helpful comments on the manuscript. This work was supported by a UCLA Academic Senate Grant and a Faculty Career Development Grant to LS. DW was supported by an NSF IGERT training grant (DGE 9972802) and a UCLA Graduate Division fellowship. UB was supported by the David and Lucille Packard Foundation as well as by the Moore Foundation.

Commercial relationships: none.
Corresponding author: Ladan Shams.
Email: ladan@psych.ucla.edu.
Address: UCLA Department of Psychology, Los Angeles, CA 90095-1563, USA.

References

Alais, D., & Burr, D. (2004). The ventriloquist effect results from near-optimal bimodal integration. *Current Biology, 14,* 257–262. [PubMed] [Article]

Bresciani, J. P., Dammeier, F., & Ernst, M. O. (2006). Vision and touch are automatically integrated for the perception of sequences of events. *Journal of Vision, 6*(5):2, 554–564, http://journalofvision.org/6/5/2/, doi:10.1167/6.5.2. [PubMed] [Article]

Bresciani, J. P., Ernst, M. O., Drewing, K., Bouyer, G., Maury, V., & Kheddar, A. (2005). Feeling what you hear: Auditory signals can modulate tactile tap perception. *Experimental Brain Research, 162,* 172–180. [PubMed]

Bülthoff, H. H., & Mallot, H. A. (1988). Integration of depth modules: Stereo and shading. *Journal of the Optical Society of America A, Optics and Image Science, 5,* 1749–1758. [PubMed]

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical-theoretic approach*. New York: Springer.

Dehaene, S. (1992). Varieties of numerical abilities. *Cognition, 44,* 1–42. [PubMed]

Dehaene, S., & Changeux, J. P. (1993). Development of elementary numerical abilities: A neuronal model. *Journal of Cognitive Neuroscience, 5,* 390–407.

Dehaene, S., Dehaene-Lambertz, G., & Cohen, L. (1998). Abstract representations of numbers in the animal and human brain. *Trends in Neurosciences, 21,* 355–361. [PubMed]

Ernst, M. O. (2005). A Bayesian view on multimodal cue integration. In G. Knoblich, I. M. Thornton, M. Grosjean, & M. Shiffrar (Eds.), *Perception of the human body from the inside out* (pp. 105–131). New York: Oxford University Press.

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

Gallistel, C. R., & Gelman, I. I. (2000). Non-verbal numerical cognition: From reals to integers. *Trends in Cognitive Sciences, 4,* 59–65. [PubMed]

Ghahramani, Z. (1995). *Computation and psychophysics of sensorimotor integration*. Cambridge: Massachusetts Institute of Technology.

Ghahramani, Z., Wolpert, D. M., & Jordan, M. I. (1997). Computational models of sensorimotor integration. In P. G. Morasso & V. Sanguineti (Eds.), *Self-organization, computational maps, and motor control* (pp. 117–147). Amsterdam, North-Holland: Elsevier.

Hillis, J. M., Ernst, M. O., Banks, M. S., & Landy, M. S. (2002). Combining sensory information: Mandatory fusion within, but not between, senses. *Science, 298,* 1627–1630. [PubMed]

Hötting, K., & Röder, B. (2004). Hearing cheats touch, but less in congenitally blind than in sighted individuals. *Psychological Science, 15,* 60–64. [PubMed]

Jacobs, R. A. (1999). Optimal integration of texture and motion cues to depth. *Vision Research, 39,* 3621–3629. [PubMed]

Körding, K. P., Beierholm, U., Ma, W. J., Quartz, S., Tenenbaum, J. B., & Shams, L. (2007). Causal inference in multisensory perception. *PLoS ONE, 2,* e943. [PubMed] [Article]

Kschischang, F. R., Frey, B. J., & Loeliger, H.-A. (2001). Factor graphs and the sum-product algorithm. *IEEE Transactions on Information Theory, 47,* 498–519.

Meck, W. H., & Church, R. M. (1983). A mode control model of counting and timing processes. *Journal of Experimental Psychology: Animal Behavior Processes, 9,* 320–334. [PubMed]

Moyer, R. S., & Landauer, T. K. (1967). Time required for judgements of numerical inequality. *Nature, 215,* 1519–1520. [PubMed]

Nieder, A., Freedman, D. J., & Miller, E. K. (2002). Representation of the quantity of visual items in the primate prefrontal cortex. *Science, 297,* 1708–1711. [PubMed]

Nieder, A., & Miller, E. K. (2004). Analogue numerical representation in Rhesus monkeys: Evidence for parallel processing. *Journal of Cognitive Neuroscience, 16,* 889–901.

Restle, F. (1970). Speed of adding and comparing numbers. *Journal of Experimental Psychology, 83,* 274–278.

Roach, N. W., Heron, J., & McGraw, P. V. (2006). Resolving multisensory conflict: A strategy for balancing the costs and benefits of audio-visual integration. *Proceedings of the Royal Society B: Biological Sciences, 273,* 2159–2168. [PubMed] [Article]

Sanabria, D., Soto-Faraco, S., & Spence, C. (2005). Assessing the effect of visual and tactile distractors on the perception of auditory apparent motion. *Experimental Brain Research, 166,* 548–558. [PubMed]

Shams, L., Kamitani, Y., & Shimojo, S. (2000). Illusions. What you see is what you hear. *Nature, 408,* 788. [PubMed]

Shams, L., Kamitani, Y., & Shimojo, S. (2002). Visual illusion induced by sound. *Cognitive Brain Research, 14,* 147–152. [PubMed]

Shams, L., Ma, W. J., & Beierholm, U. (2005). Sound-induced flash illusion as an optimal percept. *Neuroreport, 16,* 1923–1927. [PubMed]

Smith, J. E. (1982). Simple algorithms for M-alternative forced-choice calculations. *Perception & Psychophysics, 31,* 95–96. [PubMed]

Stocker, A. A., & Simoncelli, E. P. (2006) Noise characteristics and prior expectations in human visual speed perception. *Nature Neuroscience, 9,* 578–585. [PubMed]

Thompson, R. F., Mayers, K. S., Robertson, R. T., & Patterson, C. J. (1970). Number coding in association cortex of the cat. *Science, 168,* 271–273. [PubMed]

van Beers, R. J., Sittig, A. C., & Denier van der Gon, J. J. (1999). Integration of proprioceptive and visual position information: An experimentally supported model. *Journal of Neurophysiology, 81,* 1355–1364. [PubMed] [Article]

van Oeffelen, M. P., & Vos, P. G. (1982). A probabilistic model for the discrimination of visual number. *Perception & Psychophysics, 32,* 163–170. [PubMed]

Violentyev, A., Shimojo, S., & Shams, L. (2005). Touch-induced visual illusion. *Neuroreport, 16,* 1107–1110. [PubMed]

Yuille, A. L., & Bülthoff, H. H. (1996). Bayesian decision theory and psychophysics. In D. C. Knill & W. Richards (Eds.), *Perception as Bayesian inference* (pp. 123–161). New York: Cambridge University Press.