Vision and touch are automatically integrated for the perception of sequences of events

Jean-Pierre Bresciani Max Planck Institute for Biological Cybernetics, Tübingen, Germany

Franziska Dammeier Max Planck Institute for Biological Cybernetics, Tübingen, Germany

Marc O. Ernst Max Planck Institute for Biological Cybernetics, Tübingen, Germany

The purpose of the present experiment was to investigate the integration of sequences of visual and tactile events. Subjects were presented with sequences of visual flashes and tactile taps simultaneously and instructed to count either the flashes (Session 1) or the taps (Session 2). The number of flashes could differ from the number of taps by ±1. For both sessions, the perceived number of events was significantly influenced by the number of events presented in the task-irrelevant modality. Touch had a stronger influence on vision than vision on touch. Interestingly, touch was the more reliable of the two modalities—less variable estimates when presented alone. For both sessions, the perceptual estimates were less variable when stimuli were presented in both modalities than when the task-relevant modality was presented alone. These results indicate that even when one signal is explicitly task irrelevant, sensory information tends to be automatically integrated across modalities. They also suggest that the relative weight of each sensory channel in the integration process depends on its relative reliability. The results are described using a Bayesian probabilistic model for multimodal integration that accounts for the coupling between the sensory estimates.

Keywords: tactile, visual, multimodal interaction, illusions, sensory systems, Bayesian integration

Introduction

For most of our interactions with the environment, several sensory channels simultaneously inform us about the same event or physical property of the object(s) we are interacting with. For instance, when typing, we simultaneously feel, see, and hear the contact of the fingers with the keys. The central nervous system (CNS) integrates the different inputs and coregisters those that are likely to be generated by the same external event to come up with a unique coherent percept (for a review, see De Gelder & Bertelson, 2003; Ernst & Bülthoff, 2004). However, sometimes, incongruent inputs can also be automatically coregistered and perceptual biases occur (Bermant & Welch, 1976; Bertelson & Radeau, 1981; Fendrich & Corballis, 2001; Guest & Spence, 2003; Jousmäki & Hari, 1998; Kitagawa & Ichihara, 2002; Morein-Zamir, Soto-Faraco, & Kingstone, 2003; Shams, Kamitani, & Shimojo, 2000; Violentyev, Shimojo, & Shams, 2005). Interestingly, several studies assessed a possible mutual influence of two sensory channels to perform the same task and only observed a one-way bias (Bermant & Welch, 1976; Guest & Spence, 2003; Kitagawa & Ichihara, 2002; Pick, Warren, & Hay, 1969; Recanzone, 2003; Shipley, 1964). In these studies, if subjects' perception focusing on Channel A was biased by a task-irrelevant signal provided by Channel B, then reversing the roles in the same task—Channel B becoming the focal channel (to attend to) and Channel A the one providing a task-irrelevant background signal—generally failed to induce any perceptive bias. For instance, Guest and Spence (2003) showed that touch can bias visual perception of roughness, but they failed to observe any effect of vision on tactile perception. This pattern of results suggests a winner-take-all integration in which the most appropriate channel fully dominates less appropriate ones (for a review, see Welch & Warren, 1980). However, a growing body of evidence tends to demonstrate that for multimodal estimates, the different sensory signals are integrated in a weighted fashion, in which the relative weight allocated to each channel is inversely proportional to its relative variance (for a review, see Ernst & Bülthoff, 2004). We hypothesized that weighted integration could also underlie perceptual estimates for which two sensory signals are available but only one is task relevant. The present experiment tested this hypothesis for the perception of sequences of visual and tactile events. Our rationale was that the winner-take-all and weighted integration models make different predictions concerning the pattern of results to be expected.

Because it is binary (i.e., all or nothing), the winner-take-all model predicts that a sensory estimate can be biased by a background signal only if the background signal is more appropriate (i.e., provides more accurate information) than the focal one. This precludes any two-way bias between two

doi: 10.1167/6.5.2 Received August 28, 2005; published April 20, 2006 ISSN 1534-7362 © ARVO

channels. In contrast, weighted integration entails a two-way bias. More specifically, weighted integration states that the weight of each signal is proportional to its reliability—reliability = 1 / variance ($r_i = 1 / \sigma_i^2$). Under the constraint that the weights sum to 1 and that the noise of the signals is Gaussian distributed and independent, these weights can be expressed as

$$w_i = r_i \Big/ \sum_j r_j. \quad (1)$$

The weighted integration model therefore predicts that the more reliable of the signals should have a stronger biasing effect when presented as background signal and to be less susceptible to bias when constituting the focal signal. If two channels have approximately the same intrinsic reliability for a given estimate, weighted integration mechanisms should give rise to a two-way (mutual) biasing influence. For large reliability differences, however (i.e., one of the channels being by far more reliable than the other for a given estimate), the pattern of responses resulting from a weighted integration could be similar to a winner-take-all situation. In such a case, it might be difficult to distinguish the two models. In line with this, the reported one-way biases (Bermant & Welch, 1976; Guest & Spence, 2003; Kitagawa & Ichihara, 2002; Pick et al., 1969; Recanzone, 2003; Shipley, 1964) could result from large reliability differences between the presented signals.

The second difference between the two models relates to the variability of the estimates. If the sensory signals are integrated in a winner-take-all manner, the variability of cross-modal estimates cannot be lower than the variability of the estimates based on the more reliable signal alone. In contrast, if the sensory signals are integrated in a weighted manner, the variability of the estimates will be smaller when the two sources of information are available than when only the focal signal is presented. This is because in that case, the perceptual estimate is based on two sources of information instead of only one, which increases the reliability (Ernst & Banks 2002; Landy, Maloney, Johnston, & Young, 1995). This increase can lead to a maximal reliability of

$$r_{\max} = \sum_i r_i. \quad (2)$$

So far, there were two attempts to model such effect on events perception using probabilistic models. One used the maximum likelihood approach (Andersen, Tiippana, & Sams, 2005); the other used one a Bayesian integration scheme (Shams, Ma, & Beierholm, 2005). Here, we investigated the interaction between vision and touch for the perception of sequences of events, and we modeled our data using the Bayesian scheme as described in Ernst (2005). The subjects were simultaneously presented with both visual flashes and tactile taps and instructed to report either the number of flashes (Session 1) or the number of taps (Session 2). The main reason why we focused on visuo-tactile interaction is that the reported susceptibilities of visual (Shams et al., 2000) and tactile perception (Bresciani et al., 2005; Hötting & Röder, 2004) to auditory-evoked bias are commensurate. Therefore, we did not expect large reliability differences between vision and touch for this kind of event-counting task. This choice enhanced the chances of observing a two-way bias if the signals are integrated in a weighted fashion.

Methods

Subjects

Ten right-handed subjects, aged 19–35 years, participated in the experiment. None of these subjects had a history of sensorimotor disorder, and all had normal or corrected-to-normal vision. All subjects gave their informed consent before taking part in the experiment, which was performed in accordance with the ethical standards laid down in the 1964 Declaration of Helsinki.

Experimental setup

The experimental setup is schematically represented in Figure 1. The subjects were seated. Their head rested on a chin and forehead rest, whereas their right forearm and hand rested palm up at belly level on a table (72 cm high) located in front of them. The visual scene was presented on a CRT monitor mounted upside down, and the subjects viewed its reflection in an opaque mirror (for a description of the apparatus, see Ernst & Banks, 2002). The visual scene consisted of a red central fixation cross (1 deg of visual angle) displayed for the whole duration of each session, and a white circle (1 deg in diameter) flashed 8.5 deg to the right of the central fixation cross during the trials. A PHANToM

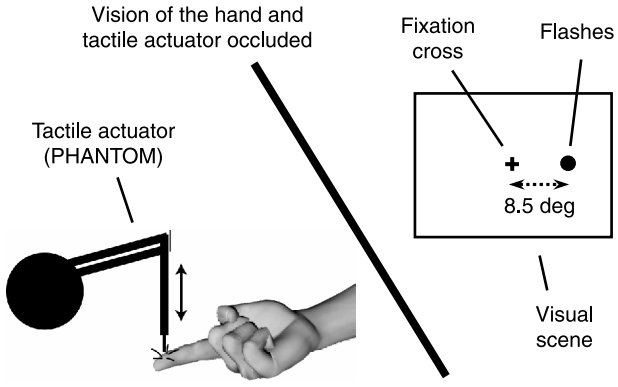

Figure 1. Experimental setup. Tactile stimulation using the PHANToM device on the left; visual scene is shown on the right. Vision of the hand and tactile stimulation device was occluded from sight.

(SensAble Technologies) force-feedback device fixed to the table was used to generate the tactile stimuli (taps of 1 N indenting subjects' skin by approximately 2 mm) via a metallic pin 3 mm in diameter. The subjects could not see their hand or the force-feedback device. Using the mirror setup, however, the sensed position of the hand corresponded to the seen position of the flashes. For the whole duration of each session, subjects wore earphones emitting a white noise (71 dB) to mask any external auditory disturbance. The subjects launched the trials and gave their responses using a keypad fixed to the left of the mirror. They were free to enter any numerical number as a response.

Procedure

In Session 1, for each trial, subjects were presented with a sequence of two to four flashes. Each flash lasted 50 ms, and the delay between the onsets of two successive flashes was 100 ms. The subjects' task was to focus on the visual sequences and to report how many flashes they saw. Tactile sequences of taps were presented simultaneously with the visual sequences. Subjects were explicitly instructed that these background tactile sequences did not relate to the focal visual sequences and were to be ignored. Four background conditions were used: "vision alone" (no tactile sequence), "one tap less" (number of taps = number of flashes − 1), "same number" (number of taps = number of flashes), and "one tap more" (number of taps = number of flashes + 1). The vision alone condition established baseline performance for visual perception. One tap less, same number, and one tap more conditions tested whether task-irrelevant tactile signals can influence visual perception. Each tap lasted 50 ms, and the delay separating the onsets of two successive taps was 100 ms. The delay between the onsets of the visual and tactile sequences was systematically adjusted so that the middle of the visual and tactile sequences coincided with respect to time (see Figure 2). This adjustment allowed a maximal overlap between the visual and tactile sequences for trials where the number of events in the respective sequences differed (i.e., one event less and one event more).

In Session 2, the design was the same as in Session 1, but taps sequences constituted the focal signal (sequences of two to four taps) and visual flashes the background signal (four conditions: tactile alone, one flash less, same number, and one flash more).

Each of the two sessions lasted 15 min, and they were performed consecutively for a total of 30 min. Five subjects, randomly assigned, started with Session 1 and five with Session 2. Each session consisted of 12 experimental conditions, combining 3 focal conditions (sequences of two, three, and four events) with 4 background conditions (focal alone [i.e., no event], one event less, same number of events, one event more). Subjects performed 10 trials per experimental condition, for a total of 120 trials per session. For each session, all 12 experimental conditions were intermixed and the trials were presented in a random order.

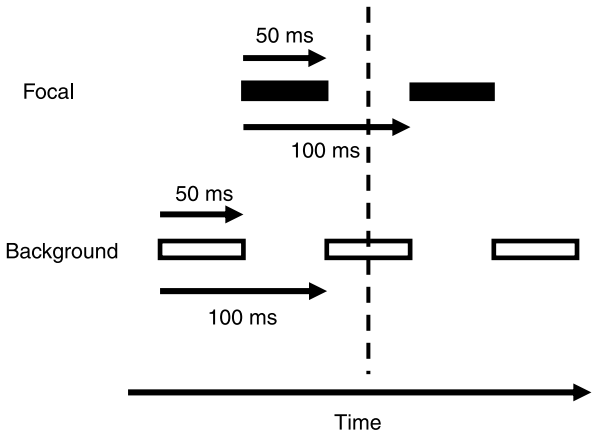

Figure 2. Temporal profiles of the stimuli in both sessions. The delay before the onset of the background sequence was systematically adjusted so that the middle of the focal and background sequences coincided with respect to time. The example given here corresponds to a trial in which two events were presented in the focal modality and one event more (i.e., three) in the background modality.

Data analysis

All statistical tests were made using repeated measures analyses of variance (ANOVAs). Post hoc comparisons using a Bonferroni adjustment for multiple comparisons ($p < .05$) were performed when necessary.

Results

We first compared the variability of responses for vision and touch. For this, we determined for each subject the distribution of responses for trials in which only the focal signal was presented and derived from there the standard deviation (σ) of responses.

As shown in Figure 3, the distribution of responses for vision alone and tactile alone was quite similar. However, the ANOVA revealed that the subjects were significantly more variable in counting the visual flashes (mean standard deviation = 0.51) than the tactile taps (mean standard deviation = 0.36), $F(1,9) = 5.96$, $p < .05$. In other words, touch is moderately more reliable than vision for this task.

Provided that visual and touch estimates are similarly reliable (touch being slightly more reliable), a weighted integration of vision and touch should give rise to a two-way biasing influence between the two modalities, with a slightly stronger effect of touch on vision than vice versa (for a more quantitative analysis, see Modeling section). In contrast, a winner-take-all integration would only give rise to a touch-evoked bias of visual perception.

To test this, we measured the influence of touch on the perceived number of flashes in Session 1 and the influence of

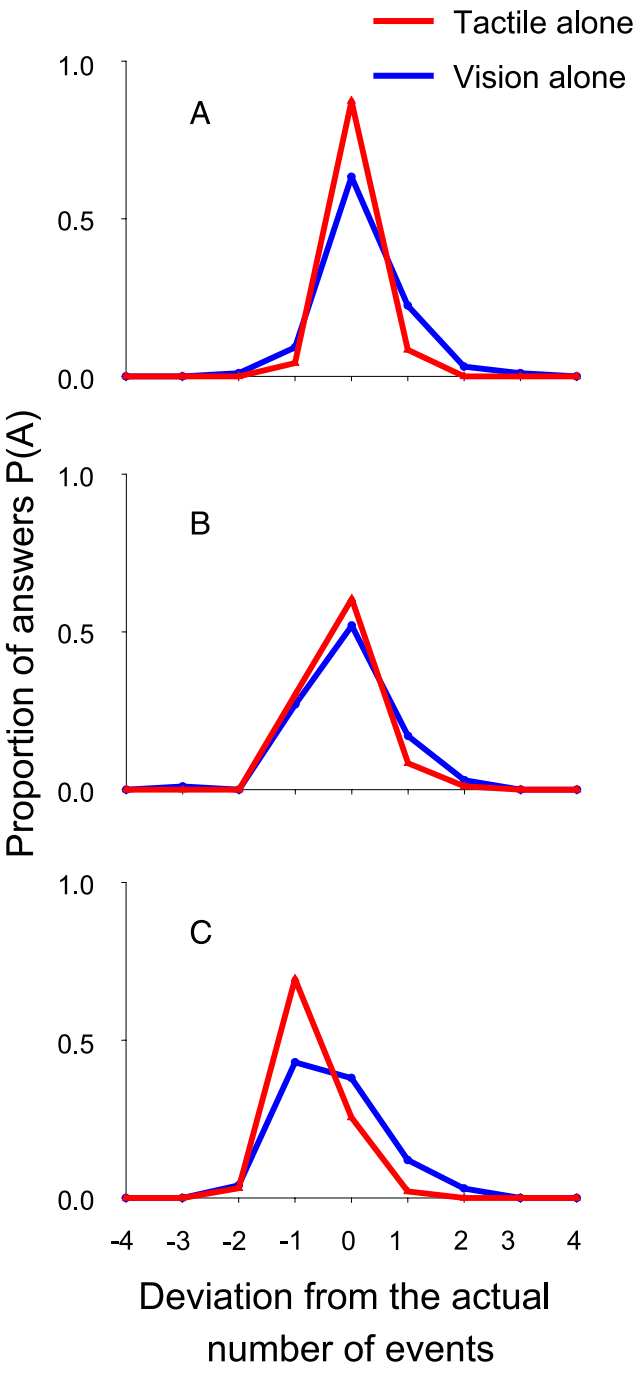

Figure 3. Average distributions of answers (percentage) for vision alone (blue lines) and tactile alone (red lines) when two events (A), three events (B), and four events (C) were presented. A deviation of zero indicates that the number of presented events was perceived correctly. Negative and positive deviations correspond to an underestimation and an overestimation of the number of events actually presented.

vision on the perceived number of taps in Session 2. For both sessions, the perceived number of events depended on the actual number of focal events, that is, the number of flashes in Session 1 and the number of taps in Session 2. More interestingly, however, in both sessions, the perceived number

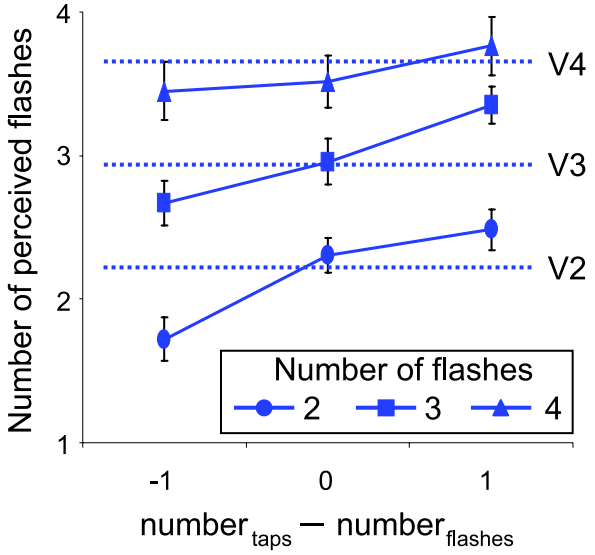

Figure 4. Number of perceived flashes as a function of both the actual number of presented flashes and the background condition. The error bars represent across-subjects standard errors. V2, V3, and V4 dotted lines represent subjects' average perception in the vision alone condition for two, three, and four flashes, respectively.

of events also depended on the background signal. In Session 1 (Figure 4), the perceived number of visual flashes was significantly influenced by the simultaneous presentation of to-be-ignored tactile taps, $F(3,27) = 25.49$, $p < .001$. Specifically, the perceived number of flashes in the one tap less condition (mean = 2.61) was significantly lower than in

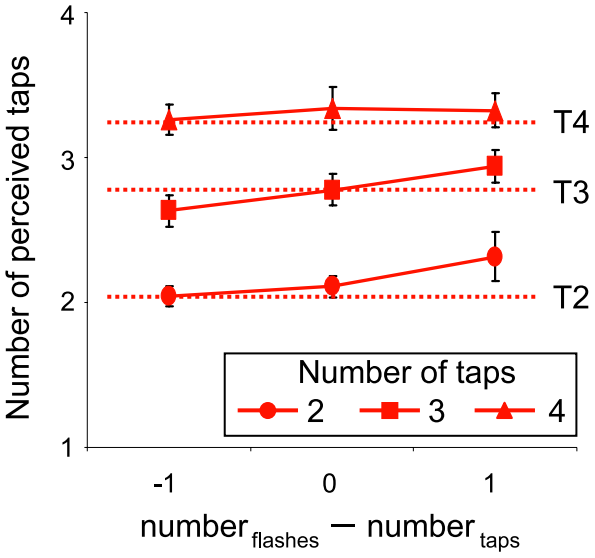

Figure 5. Number of perceived taps as a function of both the actual number of delivered taps and the background condition. The error bars represent across-subjects standard errors. T2, T3, and T4 dotted lines represent subjects' average perception in the tactile alone condition for two, three, and four taps, respectively.

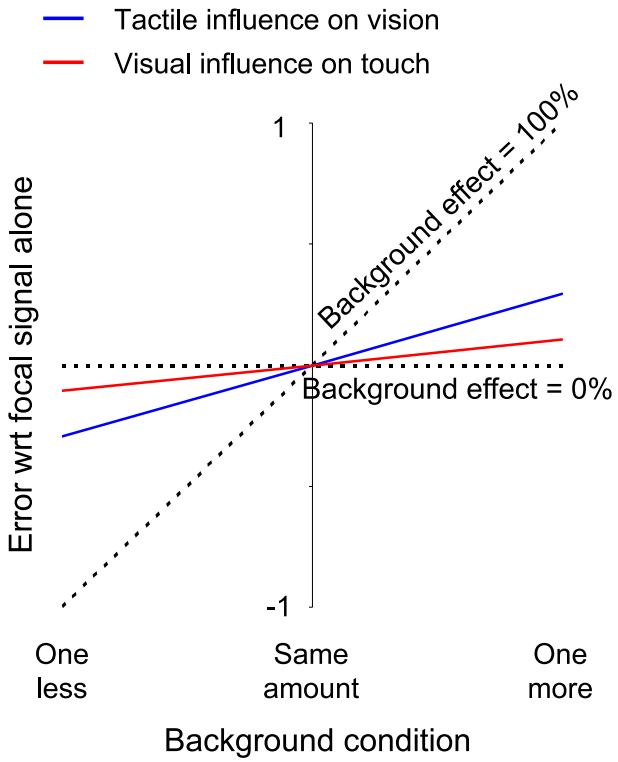

Figure 6. Regression lines showing the average effect of the background signal on subjects' perception of the focal signal. The blue line (y = 0.2928x, R^2 = .9983) represents the effect of the background tactile signal on visual perception of flashes (i.e., a tactile weight of 0.2928), whereas the red line (y = 0.1072x, R^2 = .9965) represents the effect of the background visual signal on tactile perception of taps (i.e., a visual weight of 0.1072). In both cases, background-evoked biases with respect to the condition where only the focal signal was presented were calculated. Those were averaged across the different focal conditions (two, three, and four events in the sequence), and the regression lines fitted to the means. For a better graphical representation, we only plot the slopes. The nonsignificant intercepts were 0.0214 and 0.0453 for the blue line and the red line, respectively. The two dotted lines "Background effect = 100%" and "Background effect = 0%" represent the slopes of the regression lines if the perception of the background signal was completely dominant ($y = x$) or had no effect at all (y = 0), respectively.

any other tactile condition, whereas the perceived number of flashes in the one tap more condition (mean = 3.20) was significantly higher than all but the vision alone condition. The vision alone (mean = 2.93) and the same number (mean = 2.93) conditions did not differ from one another.

Similarly, in Session 2 (Figure 5), the perceived number of taps depended on the number of simultaneously presented visual flashes, $F(3,27) = 7.58$, $p < .01$. The overall effect size, however, was smaller than in Session 1 (only the one flash less [mean = 2.64] and one flash more [mean = 2.86] conditions differed from one another).

As mentioned previously, because touch turns out to be more reliable than vision for the experimental conditions chosen, the weighted integration model predicted touch to have a stronger biasing influence on vision than vision on touch. To quantify this, we first corrected for the overall response bias. This was done by subtracting the mean of the responses obtained in the trials where only the focal signal was presented (i.e., vision alone and tactile alone conditions, respectively) from the mean of the responses for the trials in which both focal and background signals were simultaneously presented. This was done for all subjects individually and for both sessions. Figure 6 shows the overall mean of this analysis averaged across subjects and number of focal events. The corresponding individual data are presented in Figure 7.

These figures show that the effect of touch on vision was more pronounced than the effect of vision on touch. In particular, the significant interaction between the focal signal and the background condition revealed that the effect of the "one event less" and "one event more" background conditions was more pronounced when the focal signal was visual (mean background-evoked bias = −0.33 and 0.26, respectively) than when the focal signal was tactile (mean background-evoked bias = −0.06 and 0.15, respectively), $F(2,58) = 13.42$, $p < .001$. For the "same number of event" background condition, there was no difference between the two focal conditions (mean background-evoked error = 0.00 when the focal signal was visual and 0.03 when the focal signal was tactile).

The more restrictive prediction of a weighted integration model is that the variability of the estimates should be reduced when two signals are available simultaneously (i.e., focal and background) as compared with when only the focal signal is presented. To test this, for each session, we averaged for each subject the standard deviations of the three conditions for which both a focal and a background signal were presented. The resulting values were compared with the individual standard deviations of focal signal alone estimates. Before averaging, we verified that the variability of

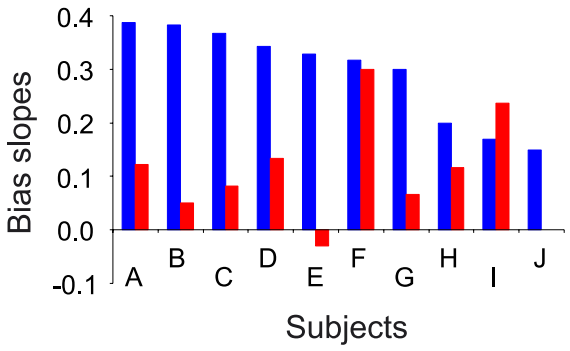

Figure 7. Effect of the background tactile signal on visual perception (blue rods) and of background visual signal on tactile perception (red rods) for each subject. A score (slope) of 1 would correspond to 100% of bias; that is, the percept is completely determined by the background signal, whereas a score of 0 would correspond to no biasing effect at all. For all subjects but one, the effect of tactile signal on visual perception is stronger than the effect of visual signal on tactile perception.

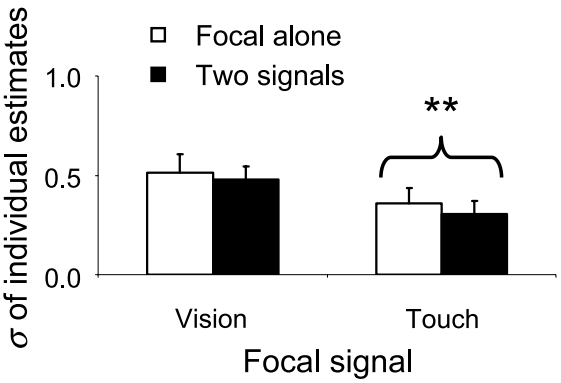

Figure 8. Average standard deviations σ of the individual estimates when only the focal signal was presented (empty columns) and when two signals were simultaneously presented (filled columns). The two columns on the left-hand side correspond to Session 1 (the task was to count the flashes), and the two columns on the right-hand side correspond to Session 2 (the task was to count the taps). The error bars represent across-subjects standard errors. When the subjects had to count the taps, being presented with two signals significantly reduced the variability of the estimates as compared with being presented with the focal signal alone.

the estimates did not vary significantly between the three conditions (number of focal events) that we averaged.

There was no effect of the number of events; thus, we averaged across the three conditions.

Figure 8 shows that in both sessions, the estimates tended to be less variable when two signals were presented than when only the focal signal was presented. This difference reached significance when the task was to count the taps (Session 2), $F(1,9) = 21.78$, $p < .01$.

Modeling

The results of the present experiment show that visual and tactile sensory signals are automatically combined for both visual (Session 1) and tactile perception (Session 2) of sequences of events. For both sessions, perceptual estimates proved to depend not only on the to-be-attended-to focal stimuli but also on simultaneously presented to-be-ignored background stimuli. In Session 1, the perceived number of visual flashes was systematically increased or decreased when more or less tactile taps were simultaneously presented. Similarly, in Session 2, the perceived number of tactile taps was modulated by the simultaneous presentation of task-irrelevant sequences of flashes. Together with the fact that bimodal stimulation has a lower variance than unimodal stimulation, these results indicate that when provided with visual and tactile signals likely originating from the same physical event, the CNS tends to automatically combine them, even when one of these signals is explicitly task irrelevant. Previous experiments using a similar paradigm demonstrated automatic combination of sensory signals (Bresciani et al., 2005; Hötting & Röder, 2004; Shams et al., 2000; Violentyev et al., 2005). The results of the present experiment extend this body of evidence. It seems therefore that automatically combining the sensory signals that are likely to be originating from the same physical event constitutes a general principle of the CNS. Such principle is functionally highly relevant if one considers that integrating multimodal signals reduces the variance of the perceptual estimates (Alais & Burr, 2004; Ernst & Banks, 2002; Gepshtein & Banks, 2003; Landy et al., 1995; Wu, Basdogan, & Srinivasan, 1999) and enhances stimulus detection (Bernstein, Clark, & Edelstein, 1969; Gielen, Schmidt, & Van den Heuvel, 1983; Hershenson, 1962; Morell, 1968; Nickerson, 1973).

Because the relative influence of one signal on the other depends on its relative reliability (Figures 6 and 7), our results seem to suggest that the perception of visual and tactile events follows the weighted integration model. If this were the case, we should be able to make more quantitative predictions using Equations 1 and 2, as this was similarly done in the recent past by several authors who have shown that the integration of multimodal signals for human perception follows the predictions of such a statistically optimal weighted integration model (e.g., Alais & Burr, 2004; Ernst & Banks, 2002; for a review, see Ernst & Bülthoff, 2004). However, as stated above, one assumption for making such quantitative prediction is that the weights of the individual signals sum up to 1. If this is the case, the weighted integration model implies complete fusion between the sensory signals. As can be seen from Figures 6 and 7, the sum of visual and tactile weights in the present experiment was less than 1. As indicated by the slope of the functions in Figure 6, the tactile signals influenced visual perception with a weight of 0.29, whereas the visual signals influenced tactile perception with a weight of only 0.11. The sum of both weights was therefore approximately 0.4 instead of 1. This means that the visual and tactile signals were not completely fused. In other words, the percept derived from the visual–tactile stimulus was not integrated into a consistent representation in which the numbers of events derived from vision and touch agree. The two modalities only biased one another. Figure 9 illustrates this concept of mutual interaction between sensory signals without the necessity of complete fusion (Figure 9C, Example 2).

Quantifying the integration process is more difficult when the sensory signals are not completely fused and only a mutual bias between the channels occurs. One way of modeling such a mutual bias, though, is to interpret the "incomplete" fusion as a coupling between the sensory channels. Such a coupling can be described in Bayesian terms using a coupling prior (for a complete description of such a model, see Ernst, 2005). By introducing this coupling prior, we add one free parameter to the model. This is the main difference with the maximum likelihood estimator (see, e.g., Ernst & Banks, 2002; Landy et al., 1995). The variance of this coupling prior, which has a Gaussian profile, determines the strength of the coupling (i.e., degree

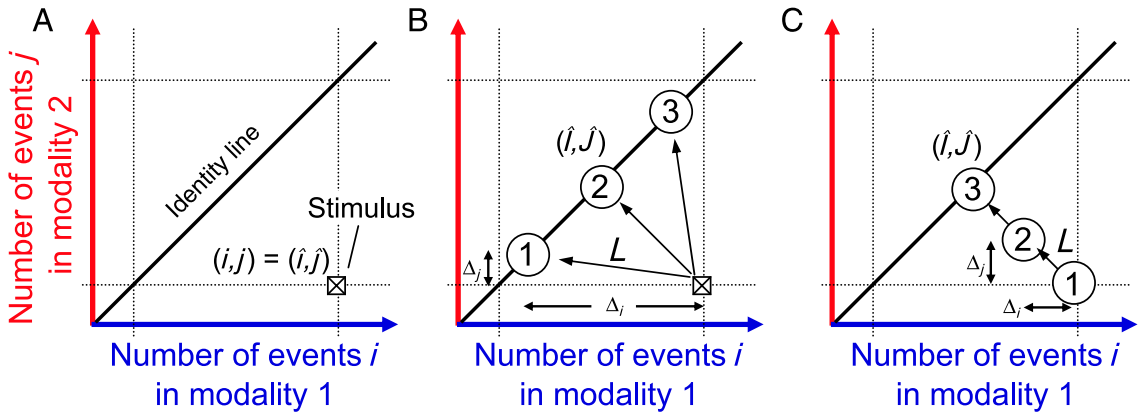

Figure 9. Examples for the interaction between sensory signals. The abscissa represents the number of events *i* of one modality, and the ordinate represents the number of events *j* of the other. (A) *i* events are presented in Modality 1 and *j* events in Modality 2; thus, the stimulus in the examples given above is (i,j) with $i > j$. If the stimuli in both modalities are presented separately (no background modality), subjects' perception for *i* events presented in Modality 1 is $\hat{\imath}$ and *j* events presented in Modality 2 is $\hat{\jmath}$. For simplicity, we assume in these examples that subjects' perception of the unimodal stimuli is unbiased; that is, $(i,j) = (\hat{\imath},\hat{\jmath})$. (B) Given are three examples (1, 2, and 3) for possible integrated percepts $(\hat{I},\hat{J})$ of the perceived number of events when both the focal and background modalities are presented simultaneously with stimulus (i,j). Example 1 demonstrates a case for which Modality 2 is dominating the percept. In Example 2, both modalities influence the percept approximately equally, and in Example 3, Modality 1 is dominating the percept. All examples fall on the diagonal line, indicating that the perceived numbers of events in both modalities agree. The difference $\Delta_i = (\hat{I} - \hat{\imath}) / (|\hat{\imath} - \hat{\jmath}|)$ represents the bias (weight) in the perception of the Modality 1 events introduced by the second modality and vice versa for $\Delta_j = (\hat{J} - \hat{\jmath}) / (|\hat{\imath} - \hat{\jmath}|)$. (C) Another set of three examples (1, 2, and 3) for possible multimodal percepts $(\hat{I},\hat{J})$. In these examples, it is not the relative influence of the two stimuli on one another that varies but the strength of the influence. The strength (the degree of coupling) can be expressed as $L = \Delta_i + \Delta_j$. A percept $(\hat{I},\hat{J})$ corresponding to a point on the identity line (Example 3) indicates complete fusion with $L = 1$. A percept $(\hat{I},\hat{J})$ that is equal to $(\hat{\imath},\hat{\jmath})$ (i.e., Example 1) indicates independence between the two modalities and $L = 0$. Percept 2 indicates an intermediate case with a mutual bias of one modality on the other ($1 > L > 0$).

of interaction) between the modalities (see Figure 10). If the variance of the prior is approaching infinity, the sources of information are independent; hence, there is no interaction between the sensory channels (i.e., they do not influence each other). In this case, the sum of the weights ($\sum_i \Delta_i$) is 0. If, on the opposite, the variance of the prior approaches 0, the sources of information are completely fused into one unified representation. A mutual influence between the sensory channels will be observed, and the weights will sum up to 1. Finally, in some intermediate cases, there is a coupling between the sensory channels but no complete fusion. Here, a mutual influence between the sensory channels can also be observed, but the sum of the weights will not sum to 1 (i.e., located between 1 and 0). Using the Bayesian approach, the percept $(\hat{I},\hat{J})$ when the stimuli in both modalities are presented simultaneously is represented by the maximum of the posterior distribution. The relative influence of one modality on the other can be determined by

$$\alpha = \arg\tan(\sigma_j^2/\sigma_i^2). \quad (3)$$

With $\alpha = 0$ deg, no influence of *j*; $\alpha = 45$ deg, equal influence of *i* and *j*; and $\alpha = 90$ deg, no influence of *i*. With the unimodal variance distributions and Equation 3, we can make a prediction for the influence of touch on vision. On average, the standard deviation for the vision-alone estimates was $\sigma_{\text{vision alone}} = 0.51$; for the touch-alone estimates, it was $\sigma_{\text{touch alone}} = 0.36$. Therefore, we predict an influence of touch on vision of $\alpha_{\text{predicted}} = 63.5$ deg. The empirical α is determined from the difference of the bimodal and unimodal percepts,

$$\alpha_{\text{empirical}} = \arg\tan(\Delta_j/\Delta_i).$$

These on average correspond to the slopes of Figure 6. Therefore, $\alpha_{\text{empirical}} = \text{argtan}(0.29 / 0.11) = 69.2$ deg, which is in good agreement with the predicted value.

The strength of coupling can be described as a weighting function between the prior and the likelihood distribution in the direction of α,

$$L = \sigma^2_{\text{likelihood}}(\alpha)/(\sigma^2_{\text{likelihood}}(\alpha) + \sigma^2_{\text{prior}}(\alpha)),$$

with

$$\sigma^2_{\text{likelihood}}(\alpha) = \sigma_i^2 \cos^2(\alpha) + \sigma_j^2 \sin^2(\alpha),$$

and

$$\sigma_{\text{prior}}(\alpha) = \sigma_p \cos(|\alpha - 45 \text{ deg}|).$$

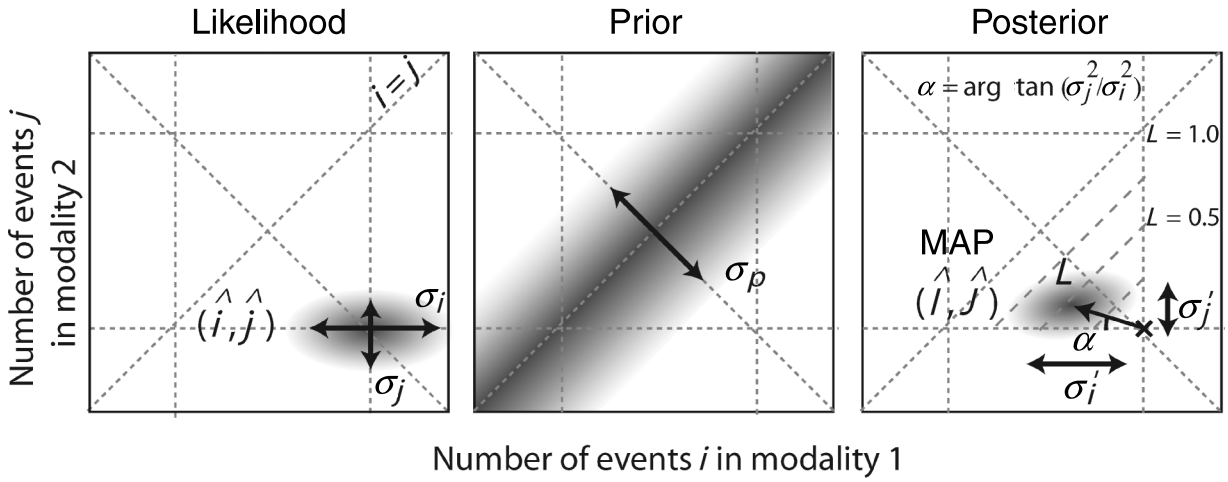

Figure 10. Bayesian model for sensory integration. The perceptual interaction between the sensory signals can be described using the Bayesian approach (Ernst, 2005). In the figure, the *likelihood* function has its maximum at $(\hat{i},\hat{j})$ and the standard deviation of the bivariate Gaussian distribution is (σ_i,σ_j). Thus, it represents the perceptual estimate when both signals are presented separately. The covariance is assumed to be 0 in this example. The *prior* represents the mapping between the signals and is thus aligned with the identity line where $i = j$. The variance of this prior (σ_p^2), which is assumed to be Gaussian distributed, represents the uncertainty in the mapping. Using Bayes' formula, the *posterior* is calculated by multiplying the likelihood with the prior distribution (and normalizing it).

The degree of coupling L as a function of σ_i, σ_j, and σ_p can be seen in Figure 11.

Such a coupling between the sensory estimates is exactly what we observed in the present experiment. As can be seen from Figure 6, the sum of the slopes was on average $L = \Delta_{\text{touch}} + \Delta_{\text{vision}} = 0.4 \pm 0.13$ (*SD* across subjects).

Knowing the individual variances (here: $\sigma_i = \sigma_{\text{vision alone}} = 0.51$, $\sigma_j = \sigma_{\text{touch alone}} = 0.36$) and the degree of coupling L, we can estimate an average variance of the coupling prior σ_p for this experiment (cf. Figure 11). This is $\sigma_p = 0.52 \pm 0.14$ (*SD* across subjects).

Associated with the coupling between the sensory estimates is a reduction in standard deviation of the combined estimate. This is indicated in Figure 10 by (σ_i',σ_j') of the posterior distribution. A simulation of the reduction in standard deviation $(\sigma_i - \sigma_i') / \sigma_i$ as a function of σ_i, σ_j, and σ_p is given in Figure 12.

As shown in Figure 12, the maximum benefits in terms of standard deviation reduction can only be expected if the sources of information are completely fused (i.e., when the variance of the prior approaches zero). A simple coupling without complete fusion can only lead to intermediate benefits. Almost no benefit is achieved with a large variance

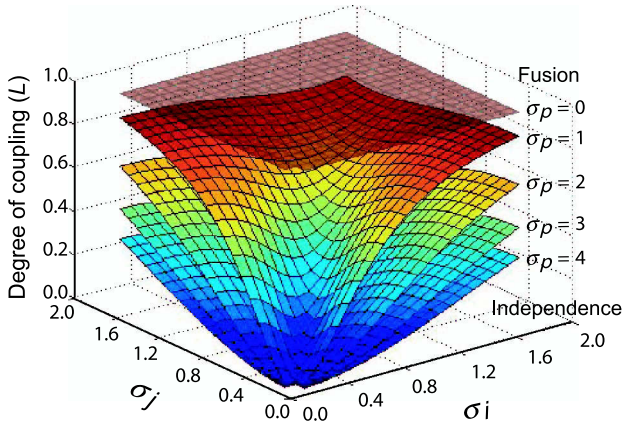

Figure 11. Degree of coupling L as a function of σ_i, σ_j, and σ_p. When $\sigma_p = 0$, there is complete fusion; thus, $L = 1$. When $\sigma_p \to \infty$, the estimates of the two signals become independent and L approaches 0.

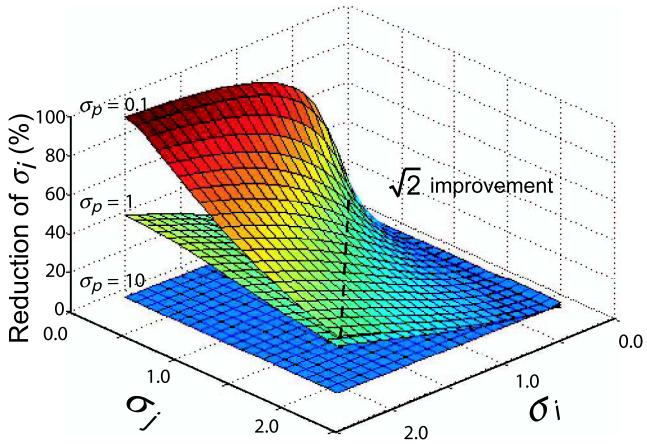

Figure 12. Reduction in standard deviation for σ_i as a function of σ_i, σ_j, and σ_p. This reduction is determined by $(\sigma_i - \sigma_i') / \sigma_i$. The maximal reduction can be achieved with a small σ_p, that is, a high degree of fusion, in case of $\sigma_i > \sigma_j$. If σ_p approaches infinity, no reduction can be gained, indicating that the sensory estimates are independent.

of the prior, which indicates independence between the sensory estimates.

In our case with a coupling prior of $\sigma_p = 0.52 \pm 0.14$ and the variability in vision ($\sigma_{\text{vision alone}} = 0.51$) and touch ($\sigma_{\text{touch alone}} = 0.36$), we predict a reduction in variability of 15.1% for vision and 7.2% for touch. Within the margin of error, this prediction agrees with the data from Figure 8, from which we can calculate an empirical reduction of 5.9% for vision and 20% for touch.

In summary, we find that when visual and haptic events are presented simultaneously, they are automatically integrated. Although the integration process does not lead to complete fusion but only to a coupling between the signals, the influence of one modality on the other is determined by the relative reliability (inverse variance) of the individual estimates. Furthermore, as a signature of sensory integration, we find the benefit of using two instead of only one modality in terms of the reduction in variance of the combined estimate. All the data can be described well using a Bayesian model with a coupling prior that determines the degree of fusion.

Discussion and conclusion

Using a paradigm in which stimuli were simultaneously presented in two sensory modalities—vision and touch—wherein only one of the modalities was task relevant, we showed that (1) the sensory modalities can mutually bias one another, the perceptual estimates being influenced by the task-irrelevant modality; (2) the influence of the more reliable modality on the less reliable one is significantly stronger than the other way around; and (3) the perceptual estimates are less variable when two modalities—focal and background—are available than when only the focal signal is presented.

These results suggest that the automatic combination of sensory signals relies on weighted integration rather than on a winner-take-all mechanism. Weighted integration is characterized by the fact that for a given estimate, the relative weight allocated to the different channels available is inversely proportional to the relative variance of each channel. It therefore predicts a mutual bias between the sensory channels, with the more reliable channel having a stronger biasing effect on and being less biased by the less reliable channel. In contrast, the winner-take-all model states that the most appropriate channel to perform the estimate fully dominates the other(s) (see Welch & Warren, 1980). It therefore only predicts an influence of the more reliable channel on the less reliable one. Previous studies based on a paradigm similar to the one used here assessed a possible two-way influence between two sensory channels (Bermant & Welch, 1976; Guest & Spence, 2003; Kitagawa & Ichihara, 2002; Pick et al., 1969; Recanzone, 2003; Shipley, 1964). These authors only observed a one-way bias. This could suggest that for this kind of bimodal task (i.e., one focal and one background modality), the sensory signals are actually integrated in a winner-take-all fashion. The results of the present experiment are inconsistent with such a hypothesis. Indeed, we observed that visual and tactile channels biased one another mutually; that is, each channel was both biased by and biased the other channel. In accordance with the predictions of a weighted integration model, the influence of the more reliable channel, namely, touch, on the less reliable channel, vision, was stronger than the influence of vision on touch. In addition, subjects' estimates were less variable when both the focal and background signals were available than when only the focal signal was presented. These results are also in agreement with a weighted integration mechanism and inconsistent with the predictions of a winner-take-all model for multisensory perception. More specifically, the winner-take-all model predicts that the lowest variability to be expected is the one observed for focal alone estimates performed with the more reliable of the two modalities, that is, the tactile modality in the current experiment. Yet, our results show that the less reliable channel reduced the variability of the more reliable one. The variability of tactile estimates was significantly lower when background visual signals were simultaneously presented. This confirms that multimodal integration occurred, thereby making use of both sensory signals available.

A matter that is very important to point out here is that the reduction in variance also rules out the possibility that the observed background-evoked bias merely resulted from the fact that the subjects sometimes focused on the focal signal and sometimes on the background signal. If that were the case, responses' variability would have been higher with both modalities than with the (less reliable of the two) focal modality only. This is because by switching, one would draw from both distributions; thus, the variance of the resulting response would be a mix between the two individual estimates' variances. Therefore, the variance of the distribution of responses to the combined stimulus could not be lower than the smaller of the two variances of the distributions of each signal alone.

There are two main methodological differences between the present experiment and previously mentioned experiments in which the integration process was quantified using the maximum likelihood estimation model (e.g., Alais & Burr, 2004; Ernst & Banks, 2002). The first is that the stimuli used here were discrete (i.e., noncontinuous). This has the implication that the distribution of responses is also discrete and so deviates from the assumption that errors are Gaussian distributed. However, as can be seen from Figure 3, the Gaussian assumption is a reasonable approximation.

Second, in our experiment, the subjects were instructed to make their estimates while focusing on one modality and ignoring the second one. Only such a method allows us to reveal whether the visual and tactile signals are completely fused or not. In spite of the incomplete fusion that we found, we were able to make clear qualitative predictions about the integration process (e.g., one-way vs. two-way bias and reduction of variance when bimodal as compared with unimodal).

Furthermore, using a free parameter, the coupling prior, we could also make quantitative predictions. More specifically, the coupling prior model allowed us to estimate the amount of bias and the reduction of variance to be expected from the integration process for an observed degree of fusion between the sensory channels. Taken together, our results suggest that vision and touch were integrated optimally in a weighted fashion according to the statistical properties of the perceptual estimates.

Acknowledgments

This work was supported by the Max-Planck Society and by the 5th Framework IST Program of the EU (IST-2001-38040, TOUCH-HapSys). We thank Roland Fleming for helpful comments on an earlier version of the manuscript.

Commercial relationships: none.
Corresponding author: Marc O. Ernst.
Email: marc.ernst@tuebingen.mpg.de.
Address: Spemannstrasse 38, 72076 Tübingen, Germany.

References

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

Andersen, T. S., Tiippana, K., & Sams, M. (2005). Maximum Likelihood Integration of rapid flashes and beeps. *Neuroscience Letters, 380,* 155–160. [PubMed]

Bermant, R. I., & Welch, R. B. (1976). Effect of degree of separation of visual–auditory stimulus and eye position upon spatial interaction of vision and audition. *Perceptual and Motor Skills, 42,* 487–493. [PubMed]

Bernstein, I. H., Clark, M. H., & Edelstein, B. A. (1969). Effects of an auditory signal on visual reaction time. *Journal of Experimental Psychology, 80,* 567–569. [PubMed]

Bertelson, P., & Radeau, M. (1981). Cross-modal bias and perceptual fusion with auditory–visual spatial discordance. *Perception & Psychophysics, 29,* 578–584. [PubMed]

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]

De Gelder, B., & Bertelson, P. (2003). Multisensory integration, perception and ecological validity. *Trends in Cognitive Science, 7,* 460–467. [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, USA: 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]

Ernst, M. O., & Bülthoff, H. H. (2004). Merging the senses into a robust percept. *Trends in Cognitive Science, 8,* 162–169. [PubMed]

Fendrich, R., & Corballis, P. M. (2001). The temporal cross-capture of audition and vision. *Perception & Psychophysics, 63,* 719–725. [PubMed] [Article]

Gepshtein, S., & Banks, M. S. (2003). Viewing geometry determines how vision and haptics combine in size perception. *Current Biology, 13,* 483–488. [PubMed] [Article]

Gielen, S. C., Schmidt, R. A., & Van den Heuvel, P. J. (1983). On the nature of intersensory facilitation of reaction time. *Perception & Psychophysics, 34,* 161–168. [PubMed]

Guest, S., & Spence, C. (2003). Tactile dominance in speeded discrimination of textures. *Experimental Brain Research, 150,* 201–207. [PubMed]

Hershenson, M. (1962). Reaction time as a measure of intersensory facilitation. *Journal of Experimental Psychology, 63,* 289–293. [PubMed]

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

Jousmäki, V., & Hari, R. (1998). Parchment-skin illusion: Sound-biased touch. *Current Biology, 8,* R190–R191. [PubMed] [Article]

Kitagawa, N., & Ichihara, S. (2002). Hearing visual motion in depth. *Nature, 416,* 172–174. [PubMed]

Landy, M. S., Maloney, L. T., Johnston, E. B., & Young, M. (1995). Measurement and modeling of depth cue combination: In defense of weak fusion. *Vision Research, 35,* 389–412. [PubMed]

Morein-Zamir, S., Soto-Faraco, S., & Kingstone, A. (2003). Auditory capture of vision: Examining temporal ventriloquism. *Cognitive Brain Research, 17,* 154–163. [PubMed]

Morell, L. K. (1968). Temporal characteristics of sensory interaction in choice reaction times. *Journal of Experimental Psychology, 77,* 14–18. [PubMed]

Nickerson, R. S. (1973). Intersensory facilitation of reaction time: Energy summation or preparation enhancement? *Psychological Review, 80,* 489–509. [PubMed]

Pick, H. L., Warren, D. H., & Hay, J. C. (1969). Sensory conflict in judgments of spatial direction. *Perception & Psychophysics, 6,* 203–205.

Recanzone, G. H. (2003). Auditory influences on visual temporal rate perception. *Journal of Neurophysiology, 89,* 1078–1093. [PubMed] [Article]

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

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

Shipley, T. (1964). Auditory flutter-driving of visual flicker. *Science, 145,* 1328–1330. [PubMed]

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

Welch, R. B., & Warren, D. H. (1980). Immediate perceptual response to intersensory discrepancy. *Psychological Bulletin, 88,* 638–667. [PubMed]

Wu, W. C., Basdogan, C., & Srinivasan, M. A. (1999). Visual, haptic, and bimodal perception of size and stiffness in virtual environments. *Proceedings of the ASME Dynamic Systems and Control Division, 67,* 19–26.