Early integration and Bayesian Causal Inference in Multisensory Perception

Ladan Shams

Department of Psychology, Department of Biomedical Engineering, Interdepartmental Neurosicence Program, University of California, Los Angeles, Franz Hall 7445B, Los Angeles, California 90095

Abstract

The last decade has witnessed great strides in understanding multisensory interactions in perception. In this chapter, we review some of the recent advances in the understanding of the neural mechanisms and computational principles of multisensory perception, with a focus on auditory-visual perception. We discuss behavioral studies demonstrating the strength of auditory influences on human visual perception, and neuroimaging studies that indicate that some of these interactions can happen at the earliest stages of visual processing in the cortex. We will describe the broad spectrum of interactions that can occur between modalities ranging from segregation to partial integration to full integration (or fusion), and review some recent computational modeling work that can account for the entire spectrum. The findings of these modeling studies suggests that the human nervous system performs inference about the causal structure of sensory signals, and for signals that are estimated to have stemmed from the same object, it integrates them to obtain a more reliable estimate of the object properties. The processes of causal inference and sensory integration are intertwined, and can be accounted well by a normative Bayesian model. Therefore, it appears that perception involves integration of multisensory signals from early stages of perceptual processing onwards, however, not all concurrent sensory signals get combined; integration vs. segregation of signals depends on a probabilistic process of inference about the causal structure of sensations.

1. Introduction

Brain function in general, and perception in particular, have been viewed as highly modular for more than a century. While phrenology is considered obsolete, its general notion of brain being

composed of compartments each devoted to a single function and independent of other functions has been the dominant paradigm especially in the context of perception (Pascual-Leone and Hamilton, 2001). In the cerebral cortex, it has been believed that the different sensory modalities are organized in separate pathways that are independent of each other, and process information almost completely in a self-contained manner until the "well-digested" processed signals converge at some higher order level of processing in the polysensory association cortical areas, wherein the unified perception of the environment is achieved. The notion of modularity of sensory modalities has been particularly strong as related to visual perception. Vision has been considered to be highly self-contained and independent of extra-modal influences. This view owes to many sources. Humans are considered to be "visual animals," and this notion has gotten underscored in the contemporary society with an ever-increasingly important role of text and images in our lives, and advent of electricity (and light at night). The notion of visual dominance has been supported by the classic and well-known studies of crossmodal interactions in which a conflict was artificially imposed between vision and another modality and found that vision overrides the conflicting sensory modality. For example, in Ventriloquist illusion, vision captures the location of discrepant auditory stimulus (Howard and Templeton, 1966). Similarly, in the "visual capture" effect, vision captures the spatial location of a tactile or proprioceptive stimulus (Rock and Victor, 1964). In McGurk effect, vision strongly and qualitatively alters the perceived syllable (McGurk and MacDonald, 1976). As a result, the influence of vision on other modalities has been acknowledged for some time. However, the influence of other modalities on vision has not been appreciated until very recently. There have been several reports of vision being influenced by another modality, however most of these have involved quantitative effects (Gebhard and Mowbray, 1959, Scheier et al., 1999, Walker and Scott, 1981, McDonald et al.,

2000, Spence and Driver, 1997, Spence et al., 1998, Stein et al., 1996). Over the last few years, two studies reported radical alterations of visual perception by auditory modality. In one case, the motion trajectory of two visual targets is some times changed from a streaming motion to a bouncing motion by a brief sound occurring at the time of visual coincidence (Sekuler et al., 1997). In this case, the motion of the visual stimuli is in principle ambiguous in the absence of sound, and one could argue that sound disambiguates this ambiguity. In another study, we found that the perceived number of pulsations of a visual flash (for which there is no obvious ambiguity) is often increased when paired with multiple beeps (Shams et al., 2000, Shams et al., 2002). This phenomenon demonstrates in an unequivocal fashion that visual perception can be altered by a non-visual signal. The effect is also very robust and resistant to changes in the shape, pattern, intensity, and timing of the visual and auditory stimuli (Shams et al., 2002, Shams et al., 2001, Watkins et al., 2006). For this reason, this illusion known as "sound-induced flash illusion" appears to reflect a mainstream mechanism of auditory-visual interaction in the brain as opposed to some aberration in neural processing. Thus, we used the sound-induced flash illusion as an experimental paradigm for investigating auditory-visual interactions in human brain.

2. Early auditory-visual interactions in the human brain

The first question we asked was at what level of processing do auditory-visual perceptual interactions occur. Do they occur at some higher-order polysensory area in the association cortex or do they involve modulation of activation along the visual cortex? We examined whether visual evoked potentials, as recorded from three electrodes in the occipitals regions of the scalp, are affected by sound. We recorded evoked potentials under visual-alone (1flash, or 2flashes), auditory-alone (2beeps), and auditory-visual (1flash2beeps) stimulus conditions. When comparing the pattern of activity associated with a second physical flash (2flash – 1flash) with

that of an illusory second flash (i.e., [1flash2beeps – 1flash – 2beeps]), we obtained a very similar temporal pattern of activity (Shams et al., 2001). Furthermore, for the 1flash2beep condition, comparing illusion and no-illusion trials revealed that the perception of illusion was associated with increased gamma-band activity in the occipital region (Bhattacharya et al., 2002). An MEG study of the flash illusion revealed modulation of activity in occipital channels by sound as early as 35-65ms post stimulus onset (Shams et al., 2005a). These results altogether indicated a mechanism of auditory-visual interaction with very short latency, and in occipital cortex. However, to map the exact location of the interactions we needed higher spatial resolution. Therefore, we performed fMRI studies of the sound-induced flash illusion. In these studies (Watkins et al., 2007, Watkins et al., 2006) the visual cortical areas were functionally mapped for each individual subject using retinotopic mapping. We contrasted auditory-visual conditions (1flash1beep, 2flash2beep) versus visual-alone conditions (1flash, 2flash). This contrast indicated auditory cortical areas, which is not surprising because in one condition there is sound and in another condition there is no sound. But interestingly, the contrast also indicated areas V1, V2 and V3, which is surprising because the visual stimulus is identical in the contrasted conditions. Therefore, these results (Watkins et al., 2006) demonstrated clearly for the first time (but see (Calvert et al., 2001)) that activity in human visual cortex as early as V1 can be modulated by non-visual stimulation. The observed increase in activation was very robust and significant. We suspected that this increase in activity may reflect a possible general arousal effect caused by sound as opposed to auditory-visual integration *per se*. Indeed attention has been previously shown to increase activity in early visual cortical areas. To address this question, we focused on the 1flash2beep condition in which some trials give rise to an illusory percept of 2 flashes (also referred to as a *fission* effect). We compared the illusion and no-illusion trials

reasoning that given that the physical stimuli are identical in both of these post-hoc-defined conditions, the arousal level should also be equal. Contrasting illusion and non-illusion trials revealed increased activity in V1 in the illusion condition (Watkins et al., 2006), indicating that perception of illusion is correlated with increased activity in V1. While this contradicts the attention hypothesis laid out earlier, one could still argue that sound may only increase arousal in some trials and those trials happen to be the illusion trials. While this argument confounds attention with integration, we could nevertheless address it using another experiment in which we included a 2flash1beep condition. On some trials of this condition, the two flashes are fused leading to an illusory percept of a single flash (also referred to as a *fusion* effect), whereas on other trials, the observers correctly perceive 2 flashes. Contrasting the illusion and non-illusion trials, we again found a significant difference in the activation level of V1, however this time the perception of sound-induced visual illusion was correlated with *decreased* activity in V1 (Watkins et al., 2007), therefore ruling out the role of attention or arousal. As mentioned above, the ERP study showed a similar temporal pattern of activity for the illusory and physical second flash. Here, we found a similar degree of V1 activation for physical and illusory double flash, and a similar degree of activation for the physical and illusory single flash (Watkins et al., 2007). These results altogether establish clearly that activity in early visual cortical areas, as early as in primary visual cortex, is modulated by sound through crossmodal integration processes.

What neural pathway could underlie these early auditory-visual interactions? Again, the last decade has witnessed overturning of another dogma; the dogma of no connectivity among the sensory cortical areas. There have been mounting evidence for direct and indirect anatomical connectivity among the sensory cortical areas (e.g., (Clavagnier et al., 2004, Falchier et al., 2002, Ghazanfar and Schroeder, 2006, Rockland and Ojima, 2003, Hackett et al., 2007)). Of particular

interest here are the findings of extensive projections from auditory core and parabelt and multisensory area STP cortical areas to V1 and V2 in monkey (Falchier et al., 2002, Rockland and Ojima, 2003, Clavagnier et al., 2004). Intriguingly, these projections appear to be only extensive for the peripheral representations in V1, and not for the foveal representations (Falchier et al., 2002). This pattern is highly consistent with the much stronger behavioral and physiological auditory modulation of vision in periphery compared to fovea that we have observed (Shams et al., 2001). (Interestingly, tactile modulation of visual processing seem to also be stronger in the periphery (Diederich and Colonius, 2007).) Therefore, it seems likely that a direct projection from A1 or a feedback projection from STS could mediate the modulations we have observed. We believe that former may be more likely, because while the activation in V1 was found to correlate with the perception of flash, the activation of area STS was always increased with perception of illusion regardless of the type of illusion (single or double-flash) (Watkins et al., 2007, Watkins et al., 2006). Therefore, these results are more readily consistent with a direct modulation of V1 projections from auditory areas.

3. Why have crossmodal interactions?

The findings discussed above as well as those discussed in other chapters, make it clear that crossmodal interactions are prevalent, and can be very strong and robust. But why? At a first glance, it may not be obvious why having crossmodal interactions would be advantageous or necessary for human's survival in the environment. Especially, in the context of visual perception, one could argue that visual perception is highly precise and accurate in so many tasks, that it may be even disadvantageous to "contaminate" it with other sensory signals that are not as reliable (which could then cause illusions or errors). Theory tells us and experimental studies have confirmed that even when a second source of information is not very reliable,

combining two sources of information can result in superior estimation compared to using only the most reliable source. Maximum likelihood estimation of an object property using two independent cues, for example an auditory estimate and a visual estimate, results in an estimate which is more reliable (more precise) than either one of the estimates. Many studies of multisensory perception have confirmed that human nervous system integrates two crossmodal estimates in a similar fashion (e.g., (Alais and Burr, 2004, Ernst and Banks, 2002, van Beers et al., 1999, Ronsse et al., 2009)). Therefore, integrating information across modalities is always beneficial. Interestingly, recent studies using single-cell recordings and behavioral measurements from macaque monkeys have provided a bridge between the behavioral manifestations of multisensory integration and neural activity, showing that the activity of multisensory (visual-vestibular) neurons is consistent with Bayesian cue integration (for a review see (Angelaki et al., 2009)).

4. The problem of causal inference

While it is beneficial to integrate information from different modalities if the signals correspond to the same object, one could see that integrating information from two different objects would not be advantageous. For example, while trying to cross the street on a foggy day, it would be beneficial to combine auditory and visual information to estimate the speed and direction of a car approaching us. It could be a fatal mistake, on the other hand, to combine the information from the sound of a car moving behind us in the opposite direction with the image of another moving car in front of us. It should be noted that humans (as with most other organisms) are constantly surrounded by multiple objects and thus multiple sources of sensory stimulation. Therefore, at any given moment, the nervous system is engaged in processing multiple sensory signals across the senses, and not all of these signals are caused by the same object, and therefore not all of

them should be bound and integrated. The problem of whether or not to combine two signals involves an (implicit or explicit) inference about whether the two signals are caused by the same object or by different objects, i.e., *causal inference*. This is not a trivial problem, and cannot be simply solved for example, based on whether the two signals originate from the same coordinate in space. The different senses have different precisions in all dimensions, including the temporal and spatial dimensions, and even if the two signals are derived from the same object/event, the noise in the environment and in the nervous system makes the sensory signals somewhat inconsistent with each other most of the time. Therefore, the nervous system needs to use as much information as possible to solve this difficult problem. It appears that whether or not two sensory signals are perceptually bound together typically depends on a combination of spatial, temporal, and structural consistency between the signals as well as the prior knowledge derived from experience about the coupling of the signals in nature. For example, moving cars often make frequency sweep sound, therefore the prior probability for combining these two stimuli should be very high. Whereas, moving cars do not typically create a bird song, therefore the prior bias for combining the image of a car and the sound of a bird is low.

Unlike the problem of causal inference in cognition that only arises intermittently, the problem of causal inference in perception has to be solved by the nervous system at any given moment, and is therefore at the heart of perceptual processing. In addition to solving the problem of causal inference, the perceptual system also needs to determine for those signals that appear to have originated from the same source, how to integrate them, i.e., to what extent, and in which direction (which modality dominating which).

5. The spectrum of multisensory combinations

To investigate these theoretical issues, we have used two complementary experimental

paradigms: a temporal numerosity judgment task (Shams et al., 2005b), and a spatial localization task (Körding et al., 2007). These two tasks are complementary in that the former is primarily a temporal task, whereas the latter is clearly a spatial task. Moreover, in the former, auditory modality dominates, whereas in the latter, vision dominates. In both of these paradigms, there are strong illusions that occur under some stimulus conditions: sound-induced flash illusion and ventriloquist illusion.

---------- Insert Figure 1 about here ----------

In the temporal numerosity experiment, a variable number of flashes were presented in the periphery simultaneously with a variable number of beeps. The task of the observers was to judge the number of flashes and beeps on each trial. In the spatial localization experiment, a Gabor patch and/or a noise burst were briefly presented at one of several locations along a horizontal line and the task of the subject was to judge the location of both the visual and auditory stimuli on each trial. In both experiments, we observed a spectrum of interactions (Fig. 1). When there was no discrepancy between the auditory and visual stimuli, the two stimuli were fused (Fig.1a left). When the discrepancy was small between the two stimuli, on a large fraction of trials they were again fused (Fig. 1a right). These trials are those wherein an illusion occurred. For example, when 1flash paired with 2 beeps was presented, on a large fraction of trials the observers reported seeing 2 flashes (sound-induced flash illusion) and hearing 2 beeps. The reverse illusion occurred when 2 flashes paired with 1 beep were seen as a single flash on a large fraction of trials. Similarly, in the localization experiment, when the spatial gap between the flash and noiseburst was small (5°), the flash captured the location of the sound on a large fraction of trials (ventriloquist illusion). In the other extreme, when the discrepancy between the

auditory and visual stimuli was large, there was little interaction if any between the two. For example, in the 1flash4beep or 4flash1beep conditions in the numerosity judgment experiment, or in the conditions where the flash was all the way to the left and noise all the way to the right or vice versa in the localization experiment, there was hardly any shift in the visual or auditory percepts relative to the unisensory conditions. We refer to this lack of interaction as segregation (Fig. 1c) because it appears that the signals are kept separate from each other. Perhaps most interestingly, in conditions in which there was a moderate discrepancy between the two stimuli, sometimes there was a partial shift of the two modalities towards each other. We refer to this phenomenon as "partial integration" (Fig. 1b). For example, in the 1flash3beep condition, sometimes the observers report seeing 2 flashes and hearing 3 beeps. Or in the condition where the flash is at -5 deg (left of fixation) and noise is at +5 degrees (right of fixation), sometimes the observers reported hearing the noise at 0 degrees and seeing the flash at -5 degrees. Therefore, in summary, in both experiments, we observed a spectrum of interactions between the two modalities. When the discrepancy is zero or small, the two modalities tend to get fused. When the conflict is moderate, partial integration may occur, and when the conflict is large, the two signals tend to be segregated (Fig. 1, right). In both experiments, the interaction between the two modalities gradually decreases as the discrepancy between the two increases (Fig. 2).

---------- Insert Figure 2 about here ----------

What would happen if we have more than two sensory signals? For example, if we have a visual, auditory, and tactile signal, as is the case most often in nature. We investigated this scenario using the numerosity judgment task (Wozny et al., 2008). We presented a variable number of

flashes paired with a variable number of beeps and a variable number of taps, providing unisensory, bisensory, and trisensory conditions pseudorandomly interleaved. The task of the participants was to judge the number of flashes, beeps and taps on each trial. This experiment provided a rich set of data that replicated the sound-induced flash illusion (Shams et al., 2000), as well as touch-induced flash illusion (Violentyev et al., 2005), as well as many previously unreported illusions. In fact, in every condition where there was a small discrepancy between two or three modalities we observed an illusion. This finding demonstrates that the interaction among these modalities is the rule rather than the exception, and the sound-induced flash illusions that have been previously reported are not "special" in the sense that they are not unusual or out of ordinary, but rather they are consistent with a general pattern of crossmodal interactions that cuts across modalities and stimulus conditions. We wondered whether these changes in perceptual reports reflect a change in response criterion as opposed to a change in perception *per se*. We calculated the sensitivity (d') change between bisensory and unisensory conditions (and between trisensory and bisensory conditions) and found statistically significant changes in sensitivity as a result of introduction of a second (or third) sensory signal in majority of cases despite the very conservative statistical criterion used. In other words, the observed illusions (both fission and fusion) reflect crossmodal integration processes, as opposed to response bias.

6. Principles governing crossmodal interactions

Is there anything surprising about the fact that there is a range of interactions between the senses? Let us examine that. Intuitively, it is reasonable for the brain to combine different sources of information to come up with the most informative guess about an object, if all the bits of information are about the same object. For example if we are holding a mug in our hand, it

makes sense that we use both haptic and visual information to estimate the shape of the mug. It is also expected for the bits of information to be fairly consistent with each other if they arise from the same object. Therefore, it would make sense for the nervous system to fuse the sensory signals when there is no or little discrepancy between the signals. Similarly, as discussed earlier, it is reasonable for the nervous system *not* to combine the bits of information if they correspond to different objects. And it is also expected for the bits of information to be highly disparate if stem from different objects. Therefore, if we are holding a mug while watching TV, it would be best *not* to combine the visual and haptic information. Therefore, segregation also makes sense from a functional point of view. How about partial integration? Is there a situation wherein partial integration would be beneficial? There is no intuitively obvious explanation for partial integration, as we do not encounter situations where two signals are only partially caused by the same object. Therefore, the phenomenon of partial integration is rather curious. Is there a single rule that can account for the entire range of crossmodal interactions including partial integration?

7. Causal inference in multisensory perception

The traditional model of cue combination (Ghahramani, 1995, Yuille and Bülthoff, 1996, Landy et al., 1995), which has been the dominant model for many years, assumes that the sensory cues all originate from the same object (Fig. 3a) and therefore, they should all be fused to obtain an optimal estimate of the object property in question. In this model, it is assumed that the sensory signals are corrupted by independent noise, and therefore are conditionally independent of each other. The optimal estimate of the source is then a linear combination of the the two sensory cues. If a Gaussian distribution is assumed for the distribution of the sensory cues, and no *a priori* bias, this linear combination would become a weighted average of the two sensory estimates, with each estimate weighted by its precision (or inverse of variance). This model has

been very successful in accounting for integration of sensory cues in various tasks and various combinations of sensory modalities (e.g., (Alais and Burr, 2004, Ernst and Banks, 2002, Ghahramani, 1995, van Beers et al., 1999)). While this model can account well for behavior when the conflict between the two signals is small (i.e., for situations of fusion, for obvious reasons), it fails to account for the rest of the spectrum (i.e., partial integration, and segregation).

---------- Insert Figure 3 about here ----------

To come up with a general model that can account for the entire range of interactions, we abandoned the assumption of a single source, and allowed each of the sensory cues to have a respective source. By allowing the two sources to be either dependent or independent, we allowed for both conditions of a common cause and conditions of independent causes for the sensory signals (Fig. 3b). We assume that the two sensory signals (x_A and x_V) are conditionally independent of each other. This follows from the assumption that up to the point where the signals get integrated, the sensory signals in different modalities are processed in separate pathways and thus are corrupted by independent noise processes. As mentioned above, this is a common assumption. The additional assumption made here is that the auditory signal is independent of the visual source (s_V) given the auditory source (s_A), and likewise for visual signal. This is based on the observation that either the two signals are caused by the same object in which case the dependence of auditory signal on the visual source is entirely captured by its dependence on the auditory source, or they are caused by different objects, in which case the auditory signal is entirely independent of the visual source (likewise for visual signal). In other words, this assumption follows from the observation that there is either a common source or

independent sources. This general model of bisensory perception (Shams et al., 2005b) results in a very simple inference rule

$$P(s_A, s_V \mid x_A, x_V) = \frac{P(x_A \mid s_A)\ P(x_V \mid s_V)\ P(s_A, s_V)}{P(x_A, x_V)} \qquad (1)$$

where the probability of the auditory and visual sources, s_A and s_V, given the sensory signals x_A and x_V is a normalized product of the auditory likelihood (i.e., the probability of getting a signal x_A given that there is a source s_A out there) and visual likelihood (i.e., the probability of getting a signal x_V given that there is a source s_V) and the prior probability of sources s_A and s_V occurring jointly. The joint prior probability $P(s_A, s_V)$ represents the implicit knowledge that the perceptual system has accumulated over the course of a life-time about the statistics of auditory-visual events in the environment. In effect, it captures the coupling between the two modalities, and therefore, how much the two modalities will interact in the process of inference. If the two signals (e.g., the number of flashes and beeps) have always been consistent in one's experience, then the expectation is that they will be highly consistent in the future, and therefore, the joint prior matrix would be diagonal (only the identical values of number of flashes and beeps are allowed, and the rest will be zero). On the other hand, if in one's experience the number of flashes and beeps are completely independent of each other, then $P(s_A, s_V)$ would be factorizable (e.g., a uniform distribution or an isotropic Gaussian distribution) indicating that the two events have nothing to do with each other, and can take on any values independtly of each other. Therefore, by having non-zero values for both $s_A = s_V$ and $s_A \neq s_V$ in this joint probability distribution, both common cause and independent cause scenarios are allowed, and the relative strength of these probabilities would determine the prior expectation of a common cause vs. independent causes. Other recent models of multisensory integration have also used joint prior to

capture the interaction between two modalities, for example, in haptic-visual numerosity judgment task (Bresciani et al., 2006) and auditory-visual rate perception (Roach et al., 2006). The model of Eq. (1) is simple, general, and readily extendable to more complex situations. For example, the inference rule for trisensory perception (Fig. 3c) would be as follows:

$$P(s_A, s_V, s_T \mid x_A, x_V, x_T) = \frac{P(x_A \mid s_A)\ P(x_V \mid s_V)\ P(x_T \mid s_T)\ P(s_A, s_V, s_T)}{P(x_A, x_V, x_T)} \qquad (2)$$

To test the trisensory perception model of Eq. (2), we modeled the three-dimensional joint prior $P(s_A, s_V, s_T)$ with a multivariate Gaussian function, and each of the likelihood functions with a univariate Gaussian function. The mean of the likelihoods were assumed to be unbiased (i.e., on average at the veridical number), and the standard deviation of the likelihoods was estimated using data from unisensory conditions. It was also assumed that the mean and variance for the prior of the three modalities is equal, and the three covariances (for three pairs of modalities) are also equal[1]. This resulted in a total of three free parameters (mean, variance, and covariance of the prior). These parameters were fitted to the data from the trisensory numerosity judgement experiment discussed earlier. The model accounted for 95% of variance in the data (676 data points) using only 3 free parameters. To test whether the three parameters rendered the model too powerful and able to account for any dataset, we scrambled the data and found that the model badly failed to account for the arbitrary data (R^2<.01). In summary, the Bayesian model of Fig.

[1] These assumptions were made in order to minimize the number of free parameters and maximize the parsimony of the model. However, the assumptions were verified by fitting a model with 9 parameters (allowing different values for the mean, variance, and covariance across modalities) to the data, and finding almost equal values for all three means, all three variances, and all three covariances.

3c can provide a remarkable account for the myriad of two-way and three-way interactions observed in the data.

8. Hierarchical Bayesian causal inference model

The model described above can account for the entire range of interactions. However, it does not directly make predictions about the perceived causal structure. In order to be able to make predictions about the perceived causal structure, one needs a hierarchical model in which there is a variable (variable C in Fig. 3d) that chooses between the different causal structures. We describe this model in the context of the spatial localization task as an example. In this model, the probability of a common cause (i.e., C=1) is simply computed using Bayes rule as follows:

$$p\left(C=1 \mid x_V, x_A\right) = \frac{p\left(x_V, x_A \mid C=1\right) p\left(C=1\right)}{p\left(x_V, x_A\right)} \qquad (3)$$

According to this rule, the probability of a common cause is simply a product of two factors. The left term in the numerator—the likelihood that the two sensory signals occur if there is a common cause—is a function of how similar the two sensory signals are. The more dissimilar the two signals, the lower this probability will be. The right term in the numerator is the *a priori* expectation of a common cause, and is a function of prior experience (how often are two signals caused by the same source in general). The denominator again is a normalization factor.
Given this probability of a common cause, the location of the auditory and visual stimulus can now be computed as follows:

$$\hat{s} = p\left(C=1 \mid x_V, x_A\right)\hat{s}_{C=1} + p\left(C=2 \mid x_V, x_A\right)\hat{s}_{C=2} \qquad (4)$$

where $\hat{s}$ denotes the overall estimate of the location of sound (or visual stimulus), and $\hat{s}_{C=1}$ and $\hat{s}_{C=2}$ denote the optimal estimates of location for the scenario of common-cause or scenario of

independent causes, respectively. The inference rule is interesting, because it is a weighted average of two optimal estimates, and it is nonlinear in x_A and x_V.

What does this inference rule mean? Let us focus on auditory estimation of location for example, and assume Gaussian functions for prior and likelihood functions over space. If the task of the observer is to judge the location of sound, then if the observer knows for certain that the auditory and visual stimuli were caused by two independent sources (e.g., a puppeteer talking, and a puppet moving) then, the optimal estimate of the location of sound would be entirely based on the auditory information and the prior: $\hat{s}_{A,C=2} = \frac{x_A / \sigma_A^2 + x_P / \sigma_P^2}{1/\sigma_A^2 + 1/\sigma_P^2}$ where σ_A and σ_P are the standard deviations of the auditory likelihood and the prior, respectively. On the other hand, if the observer knows for certain that the auditory and visual stimuli were caused by the same object (e.g., a puppet talking and moving), then the optimal estimate of the location of sound would take visual information into account: $\hat{s}_{A,C=1} = \frac{x_V / \sigma_V^2 + x_A / \sigma_A^2 + x_P / \sigma_P^2}{1/\sigma_V^2 + 1/\sigma_A^2 + 1/\sigma_P^2}$. In nature, the observer is hardly ever certain about the causal structure of the events in the environment, and in fact it is the job of the nervous system to solve that problem. Therefore, in general, the nervous system would have to take both of these possibilities into account, thus, the overall optimal estimate of the location of sound happens to be a weighted average of the two optimal estimates each weighted by their respective probability as in Eq. (3). It can now be understood how partial integration can result from this optimal scheme of multisensory perception.

It should be noted that Eq. (4) is derived assuming a mean squared error cost function. This is a common assumption, and roughly speaking, it means that the nervous system tries to minimize the average magnitude of error. The mean squared error function is minimized if the mean of the posterior distribution is selected as the estimate. The estimate shown in Eq. (4) corresponds to

the mean of the posterior distribution, and as it is a weighted average of the estimates of the two causal structures (i.e., $\hat{s}_{A,C=2}$ and $\hat{s}_{A,C=1}$), it is referred to as "model averaging." If, on the other hand, the goal of the perceptual system is to minimize the number of times that an error is made, then the maximum of the posterior distribution would be the optimal estimate. In this scenario, the overall estimate of location would be the estimate corresponding to the causal structure with the higher probability, and thus, this strategy is referred to as "model selection." Although the model averaging strategy of Eq. (4) provides estimates that are never entirely consistent with either one of the two possible scenarios (i.e., with what occurs in the environment), this strategy does minimize the magnitude of error on average (the mean squared error) more than any other strategy, and therefore it is optimal given the cost function.

9. Relationship with non-hierarchical causal inference model

The hierarchical causal inference model of Eq. (3) can be thought of as a special form of the non-hierarchical causal inference model of Eq. (1). By integrating out the hidden variable C, the hierarchical model can be recast as $p(s_A, s_V \mid x_A, x_V) = \frac{p(x_A \mid s_A)p(x_V \mid s_V)p(s_A, s_V)}{p(x_A, x_V)}$ where $p(s_A, s_V) = p(C=1)p(s) + p(C=2)p(s_A)p(s_V)$. In other words, the hierachical model is a special form of the non-hierarchical model in which the joint prior is a mixture of two priors, a prior corresponding to the independent sources, and a prior corresponding to common cause. The main advantage of the hierarchical model over the non-hierarchical model is that it performs causal inference explicitly and allows making direct predictions about perceived causal structure (C).

10. Hierarchical causal inference model versus human data

We tested whether the hierarchical causal inference model can account for human auditory-

visual spatial localization (Körding et al., 2007). We modeled the likelihood and prior over space using Gaussian functions. We assumed that the likelihood functions are on average centered around the veridical location. We also assumed that there is a bias for the center (straight-ahead) location. There were four free parameters that were fitted to the data: the prior probability of a common cause, the standard deviation of the visual likelihood (i.e., the visual sensory noise), the standard deviation of auditory likelihoods (i.e., the auditory sensory noise), and the standard deviation of the prior over space (i.e., the strength of the bias for center). Because the width of the Gaussian prior over space is a free parameter, if there is no such bias for center position, the parameter will take on a large value, practically rendering this distribution uniform, and thus, the bias largely non-existent.

The model accounted for 97% of variance in human observer data (1225 data points) using only 4 free parameters (Körding et al., 2007). This is a remarkable fit, and as before, is not due to the degrees of freedom of the model, as the model cannot account for arbitrary data using the same number of free parameters. Also, if we set the value of the four parameters using some common-sense values or the published data from other studies, and compare the data with the predictions of the model with no free parameters, we can still account for the data similarly well.

We tested whether model averaging (Eq. 4) or model selection (see above) explains the observers data better, and found that observers' responses were highly more consistent with model averaging than model selection.

In our spatial localization experiment, we did not ask participants to report their perceived causal structure on each trial. However, Wallace and colleagues did ask their subjects to report whether they perceive a unified source for the auditory and visual stimuli on each trial (Wallace et al., 2004). The hierarchical causal inference model can account for their published data; both for the

data on judgments of unity, and the spatial localizations and interactions between the two modalities (Körding et al., 2007).

We compared this model with other models of cue combination on the spatial localization dataset. The causal inference model accounts for the data substantially better than the traditional forced-fusion model of integration, and better than two recent models of integration that do not assume forced fusion (Körding et al., 2007). One of these models was a model developed by Bresciani et al. (2006) that assumes a Gaussian ridge distribution as the joint prior, and the other one was a model developed by Roach et al. (2006) that assumes the sum of a uniform distribution and a Gaussian ridge as the joint prior.

We tested the hierarchical causal inference model on the numerosity judgment data described earlier. The model accounts for 86% of variance in the data (576 data points) only using 4 free parameters (Beierholm, 2007). We also compared auditory-visual interactions and visual-visual interactions in the numerosity judgment task, and found that both crossmodal and within-modality interactions can be explained using the causal inference model, with the main difference between the two being in the *a priori* expectation of a common cause (i.e., pcommon). The prior probability of a common cause for visual-visual condition was higher than that of the auditory-visual condition (Beierholm, 2007). Hospedales and Vijayakumar (2009) have also recently shown that an adaptation of the causal inference model for an oddity detection task accounts well for both within-modality and crossmodal oddity detection of observers. Consistent with our results, they found the prior probability of a common cause to be higher for the within-modality task compared to the cross-modality task.

In summary, we found that the causal inference model accounts well for two complementary sets of data (spatial localization and numerosity judgment), it accounts well for data collected by

another group, it outperforms the traditional and other contemporary models of cue combination (on the tested dataset), and it provides a unifying account of within-modality and cross-modality integration.

11. Independence of priors and likelihoods

These results altogether strongly suggest that human observers are Bayes-optimal in multisensory perceptual tasks. What does it exactly mean to be Bayes-optimal? The general understanding of Bayesian inference is that inference is based on two factors, likelihood and prior. Likelihood represents the sensory noise (in the environment or in the brain), whereas prior captures the statistics of the events in the environment, and therefore, the two quantities are independent of each other. While this is the general interpretation of Bayesian inference, it is important to note, that demonstrating that observers are Bayes-optimal under one condition does not necessarily imply that the likelihoods and priors are independent of each other. It is quite possible that changing the likelihoods would result in a change in priors or vice versa. Given that we are able to estimate likelihoods and priors using the causal inference model, we can empirically investigate the question of independence of likelihoods and priors. Furthermore, it is possible that the Bayes-optimal performance is achieved without using Bayesian inference (Maloney and Mamassian, 2009). For example, it has been described that an observer using a table-look up mechanism can achieve near-optimal performance using reinforcement learning (Maloney and Mamassian, 2009). Because the Bayes-optimal performance can be achieved by using different processes, it has been argued that comparing human observer performance with a Bayesian observer in one setting alone is not sufficient as evidence for Bayesian inference as a process model of human perception. For these reasons, Maloney and Mamassian (2009) have proposed *transfer criteria* as more powerful experimental tests of Bayesian decision theory as a

process model of perception. The transfer criterion is to test whether the change in one component of decision process (i.e., likelihood, prior, decision rule) leaves the other components unchanged. The idea is that if the perceptual system indeed engages in Bayesian inference, a change in likelihoods, for example, would not affect the priors. However, if the system uses another process such as table-lookup then it would fail these kinds of transfer tests.

We asked whether priors are independent of likelihoods (Beierholm et al., 2009). To address this question we decided to induce a strong change in the likelihoods and examine whether this will lead to a change in priors. To induce a change in likelihoods, we manipulated the visual stimulus. We used the spatial localization task, and tested participants under two visual conditions, one with high-contrast visual stimulus (Gabor patch), and one with low-contrast visual stimulus. The task, procedure, auditory stimulus and all other things were identical across the two conditions which were tested in two separate sessions. The two sessions were apart by one week, so that if the observers learn the statistics of the stimuli during the first session, the effect of this learning would disappear by the time of the second session. The change in visual contrast was drastic enough to cause the performance on visual-alone trials to be lower than that of the high-contrast condition by as much as 41%. The performance on auditory-alone trials did not change significantly because the auditory stimuli were unchanged. The model accounts for both sets of data very well (R^2=.97 for high contrast, and R^2=.84 for low-contrast session). Therefore, the performance of the participants appears to be Bayes-optimal in both the high-contrast and low-contrast conditions. Considering that the performance in the two sessions was drastically different (substantially worse in the low-contrast condition), and considering that the priors are estimated from the behavioral responses, there is no reason to believe that the priors in these two sessions would be equal (as they are derived from very different sets of data). Therefore, if the

estimated priors do transpire to be equal between the two sessions, that would provide a strong evidence for independence of priors from likelihoods.

---------- Insert Figure 4 about here ----------

If the priors are equal, then swapping them between the two sessions should not hurt the goodness of fit to the data. We tested this. We used priors estimated from the low-contrast data to predict high-contrast data, and the priors estimated from the high-contrast data to predict the low-contrast data. The result was surprising: the goodness of fit remained almost as good (R^2=.97 and R^2=.81) as using priors from the same dataset (Beierholm et al., 2009). Next, we directly compared the estimated parameters of the likelihood and prior functions for the two sessions. The model was fitted to each individual subject's data, and the likelihood and prior parameters were estimated for each subject for each of the two sessions separately. Comparing the parameters across subjects (Fig. 4) revealed a statistically significant ($p<0.0005$) difference only for the visual likelihood (showing a higher degree of noise for the low-contrast condition). No other parameter (neither the auditory likelihood nor the two prior parameters) was statistically different between the two sessions. Despite a large difference between the two visual likelihoods (by more than 10 standard deviations) no change was detected in either probability of a common cause nor the prior over space. Therefore, these results suggest that priors are encoded independently of the likelihoods (Beierholm et al., 2009). These findings are consistent with the findings of a previous study showing that the change in the kind of perceptual bias transfers qualitatively to other types of stimuli (Adams et al., 2004).

12. Conclusions

Together with a wealth of other accumulating findings, our behavioral findings suggest that crossmodal interactions are ubiquitous, strong and robust in human perceptual processing. Even visual perception that has traditionally been believed to be the dominant modality and highly self-contained can be strongly and radically influenced by crossmodal stimulation. Our ERP, MEG, and fMRI findings consistently show that visual processing is affected by sound at the earliest levels of cortical processing, namely at V1. This modulation reflects a crossmodal integration phenomenon as opposed to attentional modulation. Therefore, multisensory integration can occur even at these early stages of sensory processing, in areas that have been traditionally held to be unisensory.

Crossmodal interactions depend on a number of factors, namely the temporal, spatial, and structural consistency between the stimuli. Depending on the degree of consistency between the two stimuli, a spectrum of interactions may result, ranging from complete integration, to partial integration, to complete segregation. The entire range of crossmodal interactions can be explained by a Bayesian model of causal inference wherein the inferred causal structure of the events in the environment depends on the degree of consistency between the signals as well as the prior knowledge/bias about the causal structure. Indeed given that humans are surrounded by multiple objects and hence multiple sources of sensory stimulation, the problem of causal inference is a fundamental problem at the core of perception. The nervous system appears to have implemented the optimal solution to this problem as the perception of human observers appears to be Bayes-optimal in multiple tasks, and the Bayesian causal inference model of multisensory perception presented here can account in a unified and coherent fashion for an entire range of interactions in a multitude of tasks. Not only the performance of observers

appears to be Bayes-optimal in multiple tasks, but the priors also appear to be independent of likelihoods, consistent with the notion of priors encoding the statistics of objects and events in the environment independent of sensory representations.

Figure Captions

Figure 1. The range of crossmodal interactions. The horizontal axis in these panels represents a perceptual dimension such as space, time, number, etc. The light bulb and loudspeaker icons represent the visual stimulus and the auditory stimulus, respectively. The eye and ear icons represent, the visual and auditory percepts, respectively. a) Fusion. Three examples of conditions in which fusion often occurs. Left, when the stimuli are congruent and they are veridically perceived. Middle, when the discrepancy between the auditory and visual stimuli is small, and the percept corresponds to a point in between the two stimuli. Right, when the discrepancy between the two stimuli is small, and one modality (in this example, vision) captures the other modality. b) Partial integration. Left, when the discrepancy between the two stimuli is moderate and the less reliable modality (in this example, vision) gets shifted towards the other modality but does not converge. Right, when the discrepancy is moderate and both modalities get shifted towards each other but not enough to converge. c) Segregation. When the conflict between the two stimuli is large, and the two stimuli do not affect each other.

Figure 2. Interaction between the auditory and visual modalities as a function of conflict. a) Visual bias (i.e., the influence of sound on visual perception) as a function of discrepancy between the number of flashes and beeps in the temporal numerosity judgment task. b) Auditory bias (i.e., the influence of vision on auditory perception) as a function of spatial gap between the two in the spatial localization task.

Figure 3. Generative model of different models of cue combination. a) The traditional model of cue combination, in which the two signals are assumed to be caused by one source. b) The causal

inference model of cue combination (Shams et al., 2005b), in which each signal has a respective cause, and the causes may or may not be related. c) The generalization of model in (b) to three signals (Wozny et al., 2008). d) The hierarchical causal inference model of cue combination (Körding et al., 2007). There are two explicit causal structures, one corresponding to common cause and one corresponding to independent causes, and the variable C chooses between the two.

Figure 4. Mean prior and likelihood parameter values across participants in two experimental sessions differing only in the contrast of visual stimulus (Beierholm et al., 2009). Blue and red denotes the values corresponding to the session with high-contrast and low-contrast visual stimulus, respectively. Error bars correspond to standard error of the mean.

Figure 1.

Figure 2.

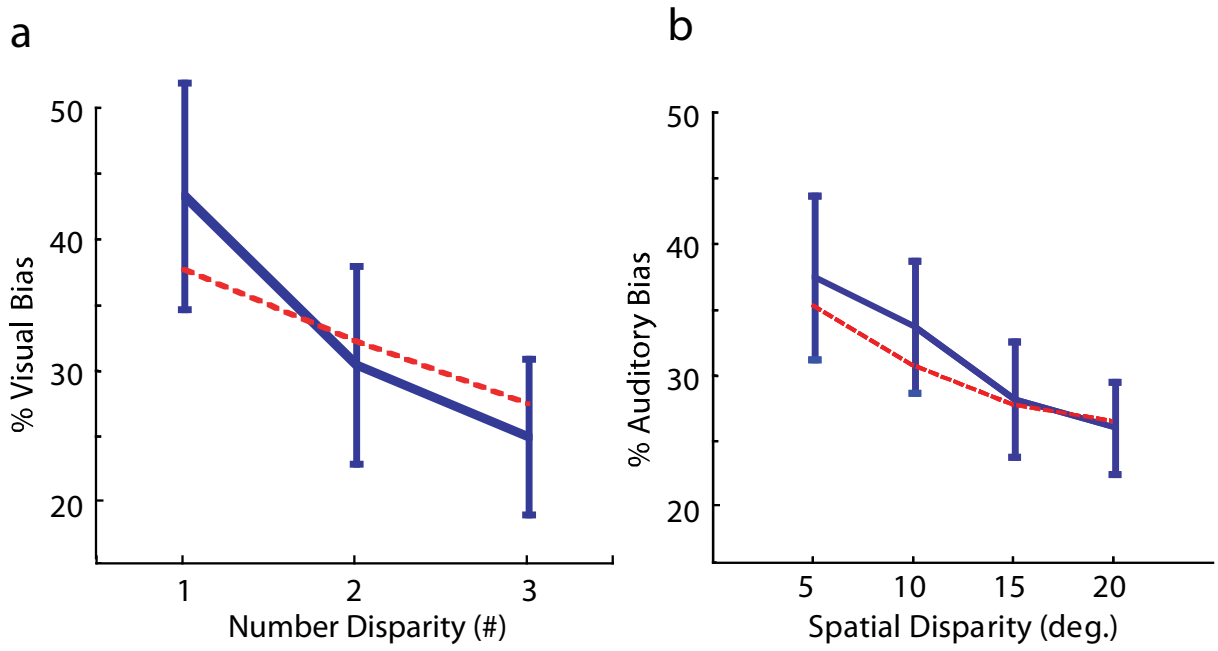
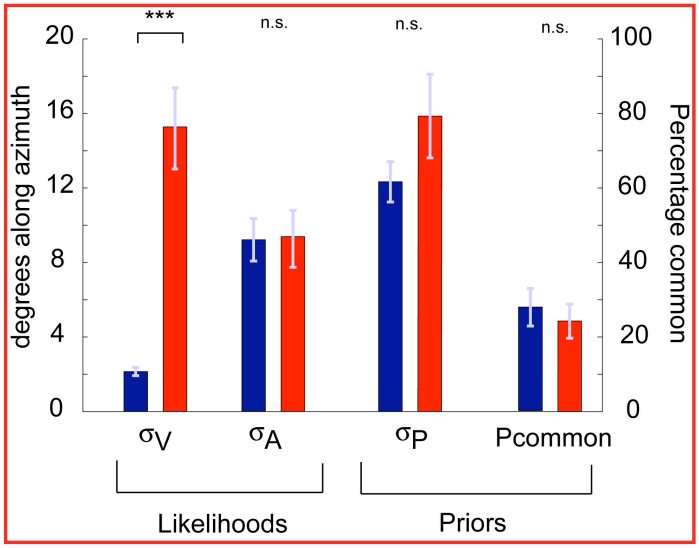

a
50
40
30
20
% Visual Bias
1
2
3
Number Disparity (#)
b
50
40
30
20
% Auditory Bias
5
10
15
20
Spatial Disparity (deg.)

Figure 3.

Figure 4.


***
n.s.
n.s.
n.s.
20
16
12
8
4
0
100
80
60
40
20
0
degrees along azimuth
Percentage common
σV
σA
σP
Pcommon
Likelihoods
Priors

References

ADAMS, W. J., GRAF, E. W. & ERNST, M. O. 2004. Experience can change the 'light-from-above' prior. *Nature Neuroscience,* 7**,** 1057-1058.

ALAIS, D. & BURR, D. 2004. The ventriloquist effect results from near-optimal bimodal integration. *Curr Biol,* 14**,** 257-62.

ANGELAKI, D. E., GU, Y. & DEANGELIS, G. C. 2009. Multisensory integration: psychophysics, neurophysiology, and computation. *Current Opinion in Neurobiology,* 19**,** 452-458.

BEIERHOLM, U. 2007. *Bayesian modeling of sensory cue combinations.* Ph.D., California Institute of Technology.

BEIERHOLM, U., QUARTZ, S. & SHAMS, L. 2009. Bayesian priors are encoded independently of likelihoods in human multisensory perception. *Journal of Vision,* 9**,** 1-9.

BHATTACHARYA, J., SHAMS, L. & SHIMOJO, S. 2002. Sound-induced illusory flash perception: role of gamma band responses. *NeuroReport,* 13**,** 1727-1730.

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**,** 554-564.

CALVERT, G., HANSEN, P. C., IVERSEN, S. D. & BRAMMER, M. J. 2001. Detection of audio-visual integration sites in humans by application of electro-physiological criteria to the BOLD effect. *Neuroimage,* 14**,** 427-438.

CLAVAGNIER, S., FALCHIER, A. & KENNEDY, H. 2004. Long-distance feedback projections to area V1: implications for multisensory integration, spatial awareness, and visual consciousness. *Cognitive Affective Behavioral Neuroscience,* 4**,** 117-126.

DIEDERICH, A. & COLONIUS, H. 2007. Modeling spatial effects in visual-tactile saccadic reaction time. *Perception & Psychophysics,* 69**,** 56-67.

ERNST, M. O. & BANKS, M. S. 2002. Humans integrate visual and haptic information in a statistically optimal fashion. *Nature,* 415**,** 429-33.

FALCHIER, A., CLAVAGNIER, S., BARONE, P. & KENNEDY, H. 2002. Anatomical evidence of multimodal integration in primate striate cortex. *Journal of Neuroscience,* 22**,** 5749-59.

GEBHARD, J. W. & MOWBRAY, G. H. 1959. On discriminating the rate of visual flicker and auditory flutter. *American Journal of Psychology,* 72**,** 521-528.

GHAHRAMANI, Z. 1995. *Computation and psychophysics of sensorimotor integration.* Ph.D. Thesis, Massachusetts Institute of Technology.

GHAZANFAR, A. & SCHROEDER, C. E. 2006. Is neocortex essentially multisensory? *Trends Cogn Sci,* 10**,** 278-285.

HACKETT, T. A., SMILEY, J. F., ULBERT, I., KARMOS, G., LAKATOS, P., DE LA MOTHE, L. A. & SCHROEDER, C. E. 2007. Sources of somatosensory input to the caudal belt areas of auditory cortex. *Perception,* 36**,** 1419-30.

HOSPEDALES, T. & VIJAYAKUMAR, S. 2009. Multisensory oddity detection as Bayesian inference. *PLoS ONE,* 4**,** e4205.

HOWARD, I. P. & TEMPLETON, W. B. 1966. *Human Spatial Orientation,* London, Wiley.

KÖRDING, K., BEIERHOLM, U., MA, W. J., TENENBAUM, J. M., QUARTZ, S. & SHAMS, L. 2007. Causal inference in multisensory perception. *PLoS ONE,* 2**,** e943.

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.

MALONEY, L. T. & MAMASSIAN, P. 2009. Bayesian decision theory as a model of human visual perception: Testing Bayesian transfer. *Visual Neuroscience,* 26**,** 147-155.

MCDONALD, J. J., TEDER-SÄLEJÄRVI, W. A. & HILLYARD, S. A. 2000. Involuntary orienting to sound improves visual perception. *Nature,* 407**,** 906-908.

MCGURK, H. & MACDONALD, J. W. 1976. Hearing lips and seeing voices. *Nature,* 264**,** 746-748.

PASCUAL-LEONE, A. & HAMILTON, R. 2001. The metamodal organization of the brain. *Prog Brain Res,* 134**,** 427-45.

ROACH, N., HERON, J. & MCGRAW, P. 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.

ROCK, I. & VICTOR, J. 1964. Vision and touch: an experimentally created conflict between the two senses. *Science,* 143**,** 594-596.

ROCKLAND, K. S. & OJIMA, H. 2003. Multisensory convergence in calcarine visual areas in macaque monkey. *International Journal of Psychophysiology,* 50**,** 19-26.

RONSSE, R., MIALL, C. & SWINNEN, S. P. 2009. Multisensory integration in dynamical behaviors: Maximum likelihood estimation across bimanual skill learning. *The Journal of Neuroscience,* 29**,** 8419-8428.

SCHEIER, C. R., NIJWAHAN, R. & SHIMOJO, S. Year. Sound alters visual temporal resolution. *In:* Investigative Ophthalmology and Visual Science, 1999 Fort Lauderdale, Florida. S4169.

SEKULER, R., SEKULER, A. B. & LAU, R. 1997. Sound alters visual motion perception. *Nature,* 385**,** 308.

SHAMS, L., IWAKI, S., CHAWLA, A. & BHATTACHARYA, J. 2005a. Early Modulation of visual cortex by sound: An MEG study. *Neuroscience Letters,* 378**,** 76-81.

SHAMS, L., KAMITANI, Y. & SHIMOJO, S. 2000. What you see is what you hear. *Nature,* 408**,** 788.

SHAMS, L., KAMITANI, Y. & SHIMOJO, S. 2002. Visual illusion induced by sound. *Cognitive Brain Research,* 14**,** 147-152.

SHAMS, L., KAMITANI, Y., THOMPSON, S. & SHIMOJO, S. 2001. Sound alters visual evoked potentials in humans. *NeuroReport,* 12**,** 3849-3852.

SHAMS, L., MA, W. J. & BEIERHOLM, U. 2005b. Sound-induced flash illusion as an optimal percept. *Neuroreport,* 16**,** 1923-1927.

SPENCE, C. & DRIVER, J. 1997. Audiovisual links in exogenous covert spatial orienting. *Perception & Psychophysics,* 59**,** 1-22.

SPENCE, C., NICHOLLS, M. E., GILLESPIE, N. & DRIVER, J. 1998. Cross-modal links in exogenous covert spatial orienting between touch, audition, and vision. *Perception and Psychophysics,* 60**,** 544-57.

STEIN, B. E., LONDON, N., WILKINSON, L. K. & PRICE, D. D. 1996. Enhancement of perceived visual intensity by auditory stimuli: A psychophysical analysis. *Journal of Cognitive Neuroscience,* 8**,** 497-506.

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.

VIOLENTYEV, A., SHIMOJO, S. & SHAMS, L. 2005. Touch-induced visual illusion. *Neuroreport,* 16**,** 1107-1110.

WALKER, J. T. & SCOTT, K. J. 1981. Auditory-Visual Conflicts in the Perceived Duration of Lights, Tones, and Gaps. *Journal of Experimental Psychology: Human Perception and Performance,* 7**,** 1327-1339.

WALLACE, M. T., ROBERSON, G. H., HAIRSTON, W. D., STEIN, B. E., VAUGHAN, J. W. & SCHIRILLO, J. A. 2004. Unifying mulitsensory signals across time and space. *Exp Brain Res,* 158**,** 252-8.

WATKINS, S., SHAMS, L., JOSEPHS, O. & REES, G. 2007. Activity in human V1 follows multisensory perception. *Neuroimage,* 37**,** 572-578.

WATKINS, S., SHAMS, L., TANAKA, S., HAYNES, J.-D. & REES, G. 2006. Sound alters activity in human V1 in association with illusory visual perception. *Neuroimage,* 31**,** 1247-1256.

WOZNY, D. R., BEIERHOLM, U. R. & SHAMS, L. 2008. Human trimodal perception follows optimal statistical inference. *Journal of Vision,* 8**,** 1-11.

YUILLE, A. L. & BÜLTHOFF, H. H. 1996. Bayesian decision theory and psychophysics. *In:* KNILL, D. C. & RICHARDS, W. (eds.) *Perception as Bayesian Inference.* Cambridge University Press.