Causal inference in multisensory perception and the brain

Dissertation

zur Erlangung des Grades eines
Doktors der Naturwissenschaften

der Mathematisch-Naturwissenschaftlichen Fakultät
und
der Medizinischen Fakultät
der Eberhard-Karls-Universität Tübingen

vorgelegt
von

Tim Rohe

aus Siegburg

April 2014

Tag der mündlichen Prüfung:	10. 10. 2014
Dekan der Math.-Nat. Fakultät:	Prof. Dr. W. Rosenstiel
Dekan der Medizinischen Fakultät:	Prof. Dr. I. B. Autenrieth
1. Berichterstatter:	Prof. Dr. U. Noppeney
2. Berichterstatter:	Prof. Dr. M. Giese
3. Berichterstatter:	Prof. Dr. J.-D. Haynes
Prüfungskommission:	Prof. Dr. U. Noppeney
	Prof. Dr. A. Fallgatter
	Prof. Dr. H.-C. Nürk
	Dr. A. Bartels

I hereby declare that I have produced the work entitled: “Causal inference in multisensory perception and the brain”, submitted for the award of a doctorate, on my own (without external help), have used only the sources and aids indicated and have marked passages included from other works, whether verbatim or in content, as such. I swear upon oath that these statements are true and that I have not concealed anything. I am aware that making a false declaration under oath is punishable by a term of imprisonment of up to three years or by a fine.

Tübingen, 29. 4. 2014

Date

Signature

Acknowledgements

First and foremost, I thank Prof. Dr. U. Noppeney for the supervision and guidance of my work, the highly constructive support and cooperation during all phases of the PhD projects, her clever ideas, and her steady encouragement. Further, I thank Prof. Dr. H.H. Bülthoff for the promotion of my PhD, and the members of my advisory board, Prof. Dr. M. Giese and Prof. Dr. H.C. Nürk, for their help and constructive suggestions. I thank Mario Kleiner for technical help on programming experiments and analysis software. Further, I thank the members of the Cognitive Neuroimaging and the Reccat groups for inspiring discussions and suggestions during group meetings. Finally, I am grateful for the encouragement and emotional support by my girlfriend Inga and my parents.

Abstract

To build coherent and veridical multisensory representations of the environment, human observers consider the causal structure of multisensory signals: If they infer a common source of the signals, observers integrate them weighted by their reliability. Otherwise, they segregate the signals. Generally, observers infer a common source if the signals correspond structurally and spatiotemporally. In six projects, the current PhD thesis investigated this causal inference model with the help of audiovisual spatial signals presented to human observers in a ventriloquist paradigm.

A first psychophysical study showed that sensory reliability determines causal inference via two mechanisms: Sensory reliability modulates how observers infer the causal structure from spatial signal disparity. Further, sensory reliability determines the weight of audiovisual signals if observers integrate the signals under assumption of a common source. Using multivariate decoding of fMRI signals, three PhD projects revealed that auditory and visual cortical hierarchies jointly implement causal inference. Specific regions of the hierarchies represented constituent spatial estimates of the causal inference model. In line with this model, anterior regions of intraparietal sulcus (IPS) represent audiovisual signals dependent on visual reliability, task-relevance, and spatial disparity of the signals. However, even in case of small signal discrepancies suggesting a common source, reliability-weighting in IPS was suboptimal as compared to a Maximum Estimation Likelihood model. By temporally manipulating visual reliability, the fifth PhD project demonstrated that human observers learn sensory reliability from current and past signals in order to weight audiovisual signals, consistent with a Bayesian learner. Finally, the sixth project showed that if visual flashes were rendered unaware by continuous flash suppression, the visual bias of the perceived auditory location was strongly reduced but still significant. The reduced ventriloquist effect was presumably mediated by the drop of visual reliability accompanying perceptual unawareness.

In conclusion, the PhD thesis suggests that human observers integrate multisensory signals according to their causal structure and temporal regularity: They integrate the signals if a common source is likely by weighting them proportional to the reliability which they learnt from the signals' history. Crucially, specific regions of cortical hierarchies jointly implement these multisensory processes.

Table of contents

1 Synopsis

1.1 Multisensory perception of a multisensory world

Humans perceive their environment as a multisensory whole because their brain effortlessly integrates distinct sensory signals — at a crowded party, we see the other guests' faces, hear their voices and feel the touch of a hand shake while smelling the food of the buffet. Subjectively, we compose this complex multisensory representation of the environment with great ease, but on closer inspection the brain accomplishes an incredibly difficult feat: just having noisy signals impinging on the sensors, the brain must first of all determine the signals' causal structure (Shams and Beierholm, 2010). Only if the signals stem from a common object, the brain should integrate them to create a multisensory representation of that object, for example a speaker's voice, face and hand shake. Signals from a separate object should be segregated, for example the voice of a different speaker. Integration entails further difficulties: Sensory noise (Faisal et al., 2008) creates stochastic discrepancies between sensory estimates of a common physical property. For example, when a speaker's voice location is slightly offset from his face position due to sensory noise, the brain has to resolve the discrepancy and figure out a unique speaker location. Further difficulties arise from the fact that multisensory signals have different representational formats (Pouget et al., 2002), for example different frames of references and neural codes, and unique qualities, such as a voice's timbre. By mastering these difficulties, the brain produces qualitatively richer representations because it combines unique sensory information and, at the same time, produces more robust representations than their unimodal counterparts (Ernst and Bulthoff, 2004). Thus, a multisensory model of the environment is the basis for successfully perceiving and acting on the world.

Behaviorally, integrating multisensory signals leads to perceptual illusions if (slightly) discrepant signals are integrated into a single coherent percept. Illusionary percepts arise from a variety of multisensory signals and can be experimentally demonstrated by introducing a discrepancy between the signals: For example, observers integrate audiovisual spatial signals such that the perceived sound shifts towards a discrepant visual signal as demonstrated by ventriloquists (i.e., the ventriloquist effect) (Thomas, 1941; Jackson, 1953; Jack and Thurlow, 1973; Radeau and Bertelson, 1977). Observers integrate a spoken syllable ("ba") with a video of a mouth speaking a slightly different syllable ("ga") and, thus, perceive an intermediate syllable ("da") (i.e., the McGurk effect) (McGurk and MacDonald, 1976). If accompanied by multiple beeps, a single visual flash is perceived as multiple flashes (i.e., the double flash illusion) (Shams et al., 2000). When rubbing their own hands, observers perceive their skin like a parchment paper if this tactile sensation is experimentally combined with a rough rubbing sound (i.e., the parchment skin effect) (Jousmaki and Hari, 1998). Observers even adopt a rubber hand to their own body scheme (Botvinick and Cohen, 1998) or feel their own body located towards an avatar's position (Lenggenhager et al., 2007) if the visually presented rubber hand or the avatar are stroked in synchrony with the observers.

These diverse multisensory illusions emerge as an epiphenomenon of integrating discrepant signals: If the signals are unified into a coherent percept, the percept appears biased as compared to the unisensory signals. Beyond coherent perception, the integration offers a benefit over the unimodal signals because it enhances the robustness of the signal estimate. Thus, multisensory signals are more quickly (Hershenson, 1962; Miller, 1982; Diederich and Colonius, 2004) and accurately (McDonald et al., 2000) processed than unisensory signals.

However, such multisensory benefits only hold if the integrated information indeed arose from the same cause. If signals are misattributed to the wrong object (e.g., a voice to the wrong speaker), the object's representation is not veridical (Roach et al., 2006). Hence, which principles do human observers apply to determine whether multisensory signals should be integrated or treated independently? How do observers integrate signals if they indeed arose from the same cause? And how do neurons process multisensory signals and which brain regions are recruited by these processes?

1.2 Principles of multisensory integration

The principles of multisensory integration were first determined in psychophysical studies using perceptual illusions. To create the illusions, observers were presented with slightly discrepant signals of multiple modalities. Thus, researcher used the shift of the illusionary percept as compared to the unimodal signals, referred to as intersensory bias, to measure the signals' weighting during integration.

The modality appropriateness hypotheses

One striking observation across different illusions was that the different modalities were dominating the illusions in specific domains. For example, vision dominated audition and proprioception in spatial judgments because the perceived auditory and proprioceptive location was strongly biased towards the visual signal location (Jackson, 1953; Pick et al., 1969; Warren and Cleaves, 1971). Similarly, the seen object size dominated over the felt object size (Rock and Victor, 1964). Thus, this 'visual capture' showed that vision dominated the integration of spatial information. By contrast, 'auditory capture' was found in the temporal domain: Multiple auditory beeps perceptually multiplied a single visual flash (Shams et al., 2000), auditory beeps temporally pulled apart two visual flashes (Morein-Zamir et al., 2003) and audition dominated temporal rate perception of audiovisual oscillatory signals (Shipley, 1964; Welch et al., 1986). Hence, visual and auditory capture showed that these modalities dominated in the spatial and temporal domain, respectively. Welch and Warren (1980) explained this pattern by the 'modality appropriateness hypothesis': The modalities process signals of a certain domain with high precision and, therefore, dominate tasks involving these signals. The modality-specific precision arises from the 'hard-wired' modality-specific formats of neural coding. For example, the visual modality favors spatial processing by using a retinotopic code (Wandell et al., 2007) while the auditory modality favors temporal processing by using a spectral tonotopic code (Recanzone and Sutter, 2008).

However, the 'modality appropriateness hypothesis' could not account for the fact that the precision of a signal does not only depend on the neural representational format: Further, signal precision depends on sensory noise (Faisal et al., 2008) and physical noise which are dynamically changing. For example, in a foggy environment visual signals might provide less precise spatial information than auditory spatial signals. Similarly, background noise degrades the temporal precision of auditory signals. Indeed, earlier studies noted that audiovisual intersensory bias depends on trial-wise relative signal strength (Radeau, 1985). However, only the advent of the Bayesian perspective on perception (Yuille and Buelthoff, 1996; Knill and Pouget, 2004; Yuille and Kersten, 2006) acknowledged that in perception observers estimate physical properties from noisy, uncertain signals: Thus, the weighting of multisensory signals should optimally depend on the dynamically changing uncertainty (or its inverse, reliability) of the signals' estimates, not the precision of the modality per se.

The MLE model

From the Bayesian perspective on perception, the brain optimally infers environmental properties by combining their noisy signals and prior knowledge. If multisensory signals arose from a common source, the perspective entails that the optimal strategy is to weight the multisensory signals and their prior proportional to their relative reliability (i.e., the inverse of their sensory variance) (Ernst and Banks, 2002). For example, an observer usually estimates the location of an audiovisual object from the visual signal under normal viewing conditions, but in foggy conditions an observer rather relies on auditory spatial information. This weighting strategy is optimal because it exploits the signals' redundancy to enhance sensory reliability: The integrated signal estimate is more reliable than each of the unimodal estimates. If the prior is uninformative, the reliability-weighted average of the multisensory signals is the maximum likelihood estimator (MLE) of the environmental property and also known as a Kalman filter in optimal control theory (Kalman, 1960).

By manipulating sensory reliability in a trial-by-trial fashion, it has been elegantly shown that current reliability determines weight of multisensory (Ernst and Banks, 2002; Battaglia et al., 2003; Alais and Burr, 2004), but also unisensory (Jacobs, 1999; Knill and Saunders, 2003; Hillis et al., 2004) and even sensorimotor signals (Kording and Wolpert, 2004). For example the visual dominance in judgments of visual-haptic object size gives way to haptic dominance if the reliability of the visual signals is reduced (Ernst and Banks, 2002; Gepshtein and Banks, 2003). Thus, the combined estimate of object size yields integration benefits close to the MLE-predicted optimum. Likewise, the strong visual bias on perceived audiovisual signal location reverses into an auditory bias if the visual signals are degraded (Battaglia et al., 2003; Alais and Burr, 2004). Overall, the MLE model explains a host of multisensory illusions by reliability-weighted signal integration. Thereby, the model exceeds the notion that signal weights are fixed for a given combination of multisensory signals which was proposed by the modality appropriateness hypothesis.

To accomplish the trial-by-trial weighing of signals, the MLE model assumes that the brain does not only represent the signals' estimate per se, but also represents 'online' sensory reliability. This could be achieved by measuring the neural response to a signal across time in a sampling-based representation (Fiser et al., 2010) or a by a probabilistic population code (PPC) (Ma et al., 2006): While the peak location of a neuronal population's response profile represents a signal's estimate, the gain of the response profile represents the momentary signal's reliability. Downstream multisensory populations then accomplish reliability-weighted integration by summing over multiple unisensory PPCs. In line with the theory, neurophysiological studies in monkeys showed that neuronal populations dynamically implement reliability-weighted integration of visual-vestibular heading signals (Fetsch et al., 2012; Fetsch et al., 2013).

However, it is currently unclear whether sensory reliability is indeed estimated instantaneously from a single signal as predicted by PPC theory, or whether information on sensory reliability is integrated over time. If the brain uses a sampling-based representation in which the neurons' activity encodes samples of a signal (Fiser et al., 2010), then reliability would be computed by the variability of neural responses over time. Thus, reliability would be naturally learned over its recent history. A sampling-based representation is consistent with a Bayesian learner who updates prior knowledge of reliability obtained from past signals with reliability information obtained from incoming signals. Thus, one PhD project ("Bayesian learning of sensory reliability in multisensory perception"; chapter 6) investigated whether humans estimate sensory reliability purely from current signals, as suggested by PPC theory, or, moreover, estimate it from past signals, as suggested by a sampling-based representation and a Bayesian learner of reliability.

Further, the MLE theory crucially assumes that signals which are to be integrated arose from a common source. Otherwise, a 'forced fusion' despite independent signal sources misattributes information which is obviously sub-optimal (Roach et al., 2006). Indeed, if larger spatiotemporal signal discrepancies suggest independent signal sources, optimal reliability-weighted integration gives way to a partial segregation of multisensory signals (Gepshtein et al., 2005; Parise et al., 2012). From a Bayesian perspective, 'forced' reliability-weighted integration is only optimal under assumption of a common source, or in other words, a unitary event emanating the signals. Of course, in natural conditions in which multisensory signals from multiple events impinge on an observer, an unconditioned 'forced fusion' assumption is not ecologically valid.

The assumption of unity

It has long been known that the assumption of unitary events giving rise to multisensory signals determines the strength of multisensory integration (Welch and Warren, 1980). The assumption of unity bases on structural factors like the spatiotemporal correspondence between multisensory signals, but also depends on cognitive factors (Radeau and Bertelson, 1977). For example, if observers explicitly judge the unity of audiovisual spatial signals, the unity judgments decline with larger temporal and spatial discrepancies between the signals (Bertelson and Radeau, 1981; Lewald and Guski, 2003; Wallace et al., 2004). At the same time, spatiotemporal signal discrepancies reduce the intersensory bias indicating diminished multisensory integration (Warren and Cleaves, 1971; Jack and Thurlow, 1973; Bertelson and Radeau, 1981; Wallace et al., 2004). The assumption of unity and the intersensory bias rest upon the same signal percepts because the bias is much stronger in trials in which observers perceive unity as compared to non-unity trials (Bertelson and Radeau, 1981; Wallace et al., 2004). Moreover, cognitive factors modulate the intersensory bias, for example the knowledge of a plausible signal source: A seen puff of steam from a kettle biases localization of a whistling sound more strongly than a light bulb biases localization of a ringing bell (Jackson, 1953).

Thus, the assumption of unity determines *whether* multisensory signals are integrated or not bound together while MLE theory describes *how* multisensory signals are integrated. How could both principles of multisensory integration be rejoined?

Models of causal inference

Reliability-weighted integration breaks down if the unity of the signal is uncertain (Gepshtein et al., 2005; Parise et al., 2012). The breakdown can be modeled by joint prior distributions of multisensory signals which mediate between signal integration and segregation. These priors can have different forms, for example a prior composed of a

Gaussian modeling correlated and a uniform distribution modeling independent signals (Roach et al., 2006; Sato et al., 2007) or a Gaussian ridge along signal discrepancy (Bresciani et al., 2006; Wozny et al., 2008). In a very similar fashion, these priors model the finding that integration is likely in case of small and segregation is likely in case of large signal discrepancies (Shams and Beierholm, 2010). However, these priors were not motivated in a principled fashion, but they were rather a post-hoc choice to account for partial signal segregation found in the data. By contrast, Kording et al. (2007) treated the problems of signal integration versus segregation and, in parallel, the assumption of unity as a probabilistic inference on the uncertain causal structure of multisensory signals (cf. Fig. 1B): If a common cause of the signals (i.e., a unitary event) is likely due to small signal discrepancies, the signals are integrated weighted by their reliability. If separate causes are likely due to large discrepancies, the signals are treated independently and, hence, segregated. Thus, the probability of a common cause, which is inferred from the signals' discrepancy, adjudicates upon signal integration versus segregation to yield a final estimate of the signal. Thereby, this hierarchical Bayesian causal inference (CI) model provides a rational strategy to elegantly reconcile the question of *whether* (superordinate question: integration vs. segregation) and *how* (subordinate: reliability-weighted integration) to integrate multisensory signals.

Mathematically, the CI model combines the reliability-weighted multisensory estimate with the task-relevant unisensory estimate proportional to the probabilities of a common or separate causes, respectively (i.e., 'causal model averaging'; cf. chapter 2.3). If the probability of a common cause is one, the CI model converges to the MLE model. If the probability of a common cause is below one (i.e., the probability of separate causes is above zero), the model can explain two important findings: First, spatiotemporal signal discrepancies reduce the intersensory bias (Warren and Cleaves, 1971; Jack and Thurlow, 1973; Bertelson and Radeau, 1981; Wallace et al., 2004) and lead to violations of MLE predictions (Gepshtein et al., 2005; Parise et al., 2012) because the discrepancies reduce the probability of a common cause, and, thereby, the influence of the reliability-weighted multisensory estimate. Second, if observes selectively focus on one modality, the task-relevant of the multisensory signals has a stronger influence on the intersensory bias compared to the task-irrelevant signal (Warren, 1979; Bertelson and Radeau, 1981). If the probability of separate causes is larger than zero, averaging the causal models naturally biases the final signal estimate in direction of the task-relevant signal. Thus, distinct signal estimates emerge if observers shift their focus between the modalities.

However, two important open questions remain on the CI model: First, it is still controversial which decision strategies observers use to combine the two signal estimates under the two potential causal structures. Kording et al. (2007) proposed that observers average the estimates weighted by the probability of their causal structures (‘causal model averaging’). By contrast, Wozny et al. (2010) found that their participants stochastically selected the estimates according to this probability (‘probability matching’). Similarly, it is unclear how observers use the probability of a common cause to explicitly judge the signals’ causal structure: They could either optimally give a common-cause response if the probability is larger than 0.5 or they could sample the causal judgments stochastically as found for the signal estimates (i.e., ‘probability matching’). Second, the CI model comprises reliability-weighted integration (i.e., the MLE model) as a special case if a common cause has been inferred with certainty. Yet, it has not been shown that reliability-weighted integration is indeed specific for trials in which participants indicate a common cause. Thus, one PhD project (‘Sensory reliability shapes causal inference via two mechanisms’, chapter 2) investigated the decision strategies for implicit causal inference involved in signal estimation and for explicit causal inference involved in explicit causal judgments. Further, we tested the specificity of reliability-weighted integration if signal sources are judged as common.

1.3 Multisensory integration and attention

In every instance of time, our sensory systems are bombarded with vast amounts of information from the outside multisensory world. On the other hand, we are only able to consciously process and act on a tiny fraction of this information. Thus, attentional filter mechanism must reduce the flood of multisensory information to select a meaningful and manageable fraction of it (Itti and Koch, 2001).

Currently, it is controversial whether such attentional filters and perceptual awareness of signals influence multisensory integration or whether multisensory integration is automatic and immune to such factors (Talsma et al., 2010). After multisensory signals have been integrated, attentional selection operates on multisensory representations (Driver, 1996; Van der Burg et al., 2008), but it is unclear at which level attentional bottom-up and top-down processes encroach on the multisensory perceptual processes per se. Evidence for automatic integration comes from psychophysical studies showing that multisensory perception is independent from spatial (Bertelson et al., 2000a; Vroomen et al., 2001) and modality-specific (Helbig and Ernst, 2008) attention. Neurons of anaesthetized animals integrate multisensory signals without awareness (Kayser et al., 2007). Similarly, neglect patient localize auditory signals biased towards visual signals even though they are not aware of the visual signals (Bertelson et al., 2000b). On the other hand, multisensory phenomena like the McGurk effect require attentional resources (Alsius et al., 2005; Alsius et al., 2007). EEG and fMRI studies revealed that neural measures of multisensory integration are enhanced by attention (Fairhall and Macaluso, 2009; Donohue et al., 2011). In conclusion, the relation of attentional and multisensory processes remains controversial, and two PhD projects aimed at investigating this interface.

Implicitly, the CI model assumes an influence of modality-specific attention. In the model, the signal estimates are biases towards the task-relevant modality if the probability that the signals came from separate sources is larger than zero. In one PhD project ("To integrate, or not to integrate: Causal inference in primary sensory and association cortices during multisensory perception", chapter 4), we tested this prediction of the CI model by manipulating the task-relevance of the signals.

In a further PhD project ('The invisible ventriloquist', chapter 7), we tested whether multisensory integration necessarily requires that observers become aware of the signals. Using continuous flash suppression (Tsuchiya and Koch, 2005), we investigated whether suppressed visual signals which were not consciously perceived biased the localization of auditory signals.

1.4 The neural basis of multisensory integration

Research on how the brain integrates multisensory signals mainly focused on how single neurons process such signals and which brain regions are recruited by these processes.

Multisensory integration in single cells and cortical hierarchies

Early work on multisensory integration in single neurons focused on the superior colliculus (SC) residing in the midbrain. SC controls the change of orientation (e.g., by saccadic eye or head movements) and, therefore, needs access to information from multiple modalities (Stein and Stanford, 2008). Animals' SC neurons integrate multisensory information because they respond to combined visual, auditory and tactile stimuli with response depression or enhancement as compared to the most effective individual stimulus (Meredith and Stein, 1983). Such multisensory interactions in SC neurons are governed by three principles (Stein and Meredith, 1993): According to the spatial principle (Meredith and Stein, 1986b), multisensory enhancement occurs if the multisensory signals emerge within the crossmodally registered receptive fields of the neuron. Multisensory depression or independence occurs if one of the multisensory stimuli emerges from outside of the receptive field. According to the temporal principle, temporal discrepancies between multisensory signals decrease multisensory interactions (Meredith et al., 1987). According to the principle of inversive effectiveness, multisensory enhancement is especially pronounced for combined multisensory signals whose individual unimodal responses are weak (Meredith and Stein, 1986a). Thus, research in single neurons revealed similar principles of multisensory integration as psychophysical studies: Multisensory integration is especially strong if the signals' spatiotemporal correspondence makes a common cause likely and the multisensory benefit is strongest if several weak signals are combined.

At the cortical level, neurophysiological and neuroimaging studies revealed multisensory integration in higher association cortex such as the posterior parietal cortex (Duhamel et al., 1998; Bremmer et al., 2001; Sadaghiani et al., 2009) and superior temporal sulcus (Bruce et al., 1981; Beauchamp et al., 2004; Werner and Noppeney, 2010). At the same time, multisensory integration is not confined to the highest regions in cortical processing hierarchies, but emerges already at lower processing levels in putatively unisensory cortex (Foxe et al., 2000; Lewis and Noppeney, 2010). For example, visual signals modulate neuronal processing in auditory cortex (Kayser et al., 2007) and auditory signals modulate processing in visual cortex (Molholm et al., 2002). These low-level multisensory interactions could arise from direct connections between unisensory regions (Falchier et al., 2002), top-down feedback from higher-order multisensory regions (Macaluso and Driver, 2005) or via crossmodal thalamic input (Lakatos et al., 2007; Cappe et al., 2009). Given that multisensory integration was found in many cortical regions, it appeared as if, provocatively stated, the whole cortex might be multisensory (Ghazanfar and Schroeder, 2006). However, multisensory integration generally increases upstream the cortical hierarchies. In low-level regions, only a small percentage of neurons responds to multisensory signals (Bizley et al., 2007) while in high-level regions the majority of neurons demonstrates multisensory responses (Dahl et al., 2009).

Bridging the levels of analysis

The finding of ubiquitous multisensory integration in the cortex illustrates that the questions of *how* and *where* the brain implements multisensory processes must be posed jointly: Because large parts of the cortical hierarchy are ultimately driven or modulated by multisensory information, the crucial question is how hierarchical levels jointly implement specific multisensory processes (Driver and Noesselt, 2008).

At this point, it is important to note that multisensory processes were often differently constrained in psychophysical studies compared to neurophysiological and neuroimaging studies. The former studies derived principles from computational considerations to investigate multisensory integration, for example how an ideal observer would solve the signal estimation problem given by noisy multisensory signals (Ernst and Banks, 2002). The latter studies used neural properties to constrain models of multisensory integration, for example the problem of different reference frames in different sensory systems (Avillac et al., 2005) or enhancement versus depression of multisensory neural responses (Meredith and Stein, 1983). These approaches partially converged on similar principles, for example the finding that multisensory integration in behavior and neurons crucially depends on spatiotemporal and semantic congruence of multisensory signals. However, principles as formulated in the MLE and CI models were largely unexplored in studies with a focus on the brain until recently (Morgan et al., 2008; Fetsch et al., 2012; Helbig et al., 2012). Thus, it is important to close the gap between psychophysical and neural models of multisensory integration to get a comprehensive understanding of multisensory processes at all levels of analysis (Fetsch et al., 2013). Further, applying psychophysical models to neural data could well constrain which specific multisensory processes are implemented in different levels of cortical hierarchies.

Bridging the levels of analysis has proven difficult because often the methods applied in both fields do not directly map onto each other (e.g., multisensory response enhancement/depression as measured by firing rates, fMRI activation and event-related EEG potentials vs. intersensory bias, parameters of psychometric functions and reaction times). Thus, models of multisensory integration were partially also constrained by methodological aspects. The advent of multivariate decoding methods applied to fMRI activation patterns (Haxby et al., 2001; Kay and Gallant, 2009; Serences and Saproo, 2012) can bridge the gap between the levels of analysis. The methods allows to apply psychophysical theories of multisensory integration to fMRI activation measured along entire cortical hierarchies (chapters 3-5): Psychophysical studies investigate the mapping from a physical space of signals to the perceptual space according to well-defined models (e.g., the perceptual transformation of a two-dimensional space spanned by independent visual and auditory spatial signals according to the MLE model; cf. Fig. 1A). In a similar way, the decoding methods reconstruct a ‘neural’ space from fMRI activation patterns of specific cortical regions which can be analyzed using the same models as applied to perceptual spaces and can be compared to these.

By decoding audiovisual spatial signals from fMRI activation patterns, chapters 3-5 investigated how principles of the MLE and the CI models are implemented by visual (Mishkin et al., 1983) and auditory (Tian et al., 2001) spatial processing hierarchies. The second project (“Cortical hierarchies jointly perform Bayesian causal inference for multisensory perception”, chapter 3) investigated how different levels of these cortical hierarchies represent the constituent spatial estimates of the CI model. Chapter 4 compared the profile of neural audiovisual weighting along cortical hierarchies against the predictions of the CI model if sensory reliability, task-relevance and spatial disparity of the signals were manipulated. Finally, the fourth project (“Suboptimal reliability-weighted integration of audiovisual spatial signals in parietal cortex”, chapter 5) tested the quantitative predictions of the MLE model on reliability-weighted integration by fitting ‘neurometric’ functions to the fMRI-decoded multisensory signals.

1.5 Thesis overview

The overarching question of this PhD thesis was how human observers integrate multisensory signals given their uncertain causal structure and dynamically changing sensory reliabilities. Further, the question was how cortical hierarchies implement these processes. The MLE (Ernst and Banks, 2002) and the CI model (Kording et al., 2007) established a common framework to investigate these questions.

Methodologically, we consistently used variants of a spatial ventriloquist paradigm (Radeau and Bertelson, 1977). In the paradigm, participants were presented with audiovisual spatial signals sampled from 3-5 spatial locations on the azimuth. The auditory signal was white noise or pure tones, either convolved with head-related transfer functions and then presented via headphone or presented directly via speakers from different locations. The visual signal was a cloud of dots presented on a screen. We manipulated the signals' spatial discrepancy as well as the visual reliability by changing the variance of the cloud of dots. In chapter 7, we rendered the visual signals partially unaware by using continuous flash suppression (Tsuchiya and Koch, 2005) which requires stereoscopic presentation of the visual signals. Participants localized the auditory or the visual signals (i.e., a manipulation of the signals' task-relevance), judged whether the signals arose from common or separate sources (chapter 2) and judged whether they had perceived the visual signal (chapter 7).

Chapter 2: Sensory reliability shapes causal inference via two mechanisms

In chapter 2, we investigated two aspects of the CI model (Kording et al., 2007): First, it is unclear which decision strategies observers employ to balance the sensory estimates under the two potential causal structures of the multisensory signals (i.e., an implicit inference of common vs. separate sources when estimating signals' physical value). Further, it is unclear which decision strategies observers use to explicitly estimate the causal structure from the posterior probability of a common source. Second, it has not been empirically tested whether reliability-weighted integration is indeed limited to signals for which observers infer a common source as predicted by the hierarchy of the CI model. To this end, participants were presented with audiovisual spatial signals of varying visual reliability. In each trial, participants localized the auditory signal and judged the causal structure of the signals.

For implicit causal inference involved in auditory localization responses, a model comparison revealed that participants balanced the potential causal structures by weighting the auditory spatial estimates under the two causal structures by their probabilities ('model averaging'). Participants optimally reported a common source when the posterior probability of a common source exceeded 0.5. Visual reliability shaped the participants' causal inferences via two mechanisms: First, visual reliability sharpened the audiovisual integration window which common-source judgments formed based on signal disparity. Second, especially when participants perceived a common source, the shift of the perceived auditory location in direction of the visual signal (i.e., the ventriloquist effect) increased with higher visual reliability while localization variability decreased.

In conclusion, the study revealed that observers employ optimal decision strategies to consider the uncertainty of the signals' causal structure when they are probed in implicit as well as explicit causal inference. Further, the study revealed that the hierarchical CI model correctly predicts that reliability-weighted integration is subordinate to causal inferences which jointly depend on sensory reliability and signal disparity. Thereby, the study showed that the CI model explains multisensory phenomena beyond the MLE model by considering it as a special case.

Because the study demonstrated that the CI model accurately explains whether (i.e., causal inference) and how (i.e., reliability-weighted integration) multisensory signals are integrated at the behavioral level, the question arose how the brain represents the constituent spatial estimates of the model. The next project targeted this question.

Chapter 3: Cortical hierarchies jointly perform Bayesian causal inference for multisensory perception

To date, it is unclear how the brain represents the constituent spatial estimates of the CI model: the reliability-weighted average under assumption of a common source, the unisensory visual or auditory—task-relevant—estimates under assumption of independent sources, and the combination of both weighted by the probability of a common or independent sources, respectively. To evaluate this question, participants were presented with audiovisual spatial signals at two levels of visual reliability whilst fMRI data was collected from regions along visual (Mishkin et al., 1983) and auditory (Tian et al., 2001) spatial processing hierarchies. Participants localized either the auditory or the visual signal.

The CI model fitted the behavioral data accurately and predicted the values of its components in all experiment conditions. Using a multivariate approach to decode these CI model components from fMRI activation patterns of regions in the hierarchies, we found that unisensory areas represent their preferred component (i.e., low-level visual regions represent the unimodal visual estimate and vice versa for low-level auditory regions). Posterior regions of the intraparietal sulcus (IPS) represented the reliability-weighted average of the signals under a common-source assumption. Anterior regions of IPS represented the combination of the reliability-weighted average and the task-relevant unisensory estimate weighted by the probability of a common versus independent sources.

In conclusion, the study suggests that cortical hierarchies jointly implement causal inference in multisensory integration. A crucial future question is which regions compute the probability of a common source involved in the multisensory computations of anterior IPS.

Chapter 4: To integrate, or not to integrate: Causal inference in primary sensory and association cortices during multisensory perception

For the study in chapter 3, we chose a strictly model-based approach to investigate causal inference. The neural representations of the audiovisual signals were analyzed with regard to CI model variables constrained by the behavioral localization responses. Thus, it was unclear whether the profile of integration along visual and auditory hierarchies matched the specific predictions of the CI model if we manipulated sensory reliability, spatial discrepancy and task-relevance of the audiovisual signals. To evaluate the profile of integration, we analyzed the data from chapter 3 and used a 'model-free' index of relative weighting of the audiovisual spatial signals encoded in fMRI voxel response patterns.

Consistent with the results from chapter 3 and the predictions of the CI model, we found that specifically IPS weighted audiovisual signals proportional to sensory reliability and task-relevance. Critically, in anterior IPS the effect of task-relevance was more pronounced for large spatial discrepancy indicating a stronger segregation of the signals when independent signal sources are more likely. Low-level visual and auditory regions predominantly represented their preferred signals. However, these regions also slightly integrated the non-preferred signal especially if the spatial discrepancy was small. Further, we found that the weighting of the signals in behavioral localization responses was highly correlated with the neural weighting in anterior IPS. This indicated the behavioral relevance of the region's spatial estimates.

To sum up, the specific profile of audiovisual integration along the cortical hierarchies confirmed that higher association cortices such as anterior IPS implement causal inference. At the same time, low-level sensory regions implement multisensory processes which are governed by different principles (e.g., the spatial principle (Stein and Meredith, 1993)).

Chapter 5: Suboptimal reliability-weighted integration of audiovisual spatial signals in parietal cortex

In chapter 5, we investigated whether audiovisual spatial signals in IPS were weighted by their relative sensory reliability as quantified by the MLE model: Sensory variances measured in unimodal conditions predicted the relative weighting of the signals and the reduction of variances in bimodal conditions. To test these predictions, we presented participants with unimodal and slightly discrepant (≤ 6°) bimodal auditory and visual spatial signals at two levels of visual reliability whilst fMRI scanning.

We derived estimates of uni- and bimodal sensory variances and relative signal weights from parameters of 'neurometric' functions which we fitted to signal locations decoded from fMRI voxel response patterns. In parallel to the psychophysical results, we found that audiovisual signals were weighted proportionally to reliability in IPS, but the weighting was not optimal compared to the MLE predictions. The main reason for suboptimal weighting was that IPS weighted the signals depending on their task-relevance. This finding was surprising because the small signal discrepancy suggested a mandatory, task-independent fusion of the signals.

Thus, the study showed that the MLE model's assumption of mandatory fusion is too strict, on the psychophysical as well as the neural level. In line with the chapters 2-4, the study showed that partial signal segregation and effects of task-relevance can be better explained by the CI model.

Chapter 6: Bayesian learning of sensory reliability in multisensory perception

According to the Bayesian perspective on perception (Yuille and Buelthoff, 1996; Ernst and Banks, 2002; Knill and Pouget, 2004), observers weight prior assumptions and multisensory signals proportional to their reliability. However, it is unclear how the brain represents reliability (or its inverse, uncertainty). The theory of probabilistic population codes (Ma et al., 2006) suggests that the brain immediately represents reliability via the gain of a neuronal population's response to a signal, without the need for learning of reliability. However, if sensory reliability changes systematically as often found in natural environments, a Bayesian learner would estimate reliability by combining prior with current reliability information.

To investigate whether human observers infer visual reliability only from current, or, moreover, past visual signals, we manipulated visual reliability according to a sinusoidal or two different random-walk sequences in a ventriloquist paradigm. The results showed that the participants indeed estimated posterior visual reliability from current as well as prior visual signals consistent with a Bayesian learner: They weighted the audiovisual signals proportionally to the posterior reliability estimate when localizing the auditory signals.

Therefore, the study showed that human observers employ Bayesian inference in sensory learning when estimating posterior visual reliability as well as in perception when they localize audiovisual signals proportional to this reliability estimate.

Chapter 7: The invisible ventriloquist

Currently, it is controversial whether multisensory processes are immune to attentional processes and do not require perceptual awareness (Talsma et al., 2010). Hence, we tested whether a visual signal still attracts the perceived sound location (i.e., the ventriloquist effect) when it is rendered perceptually unaware by continuous flash suppression (CFS) (Tsuchiya and Koch, 2005).

The participants reported that CFS rendered the visual signal unaware in most trials, but sometimes observers still perceived the signal. Thus, the visual biases could be investigated in trials in which participants where either aware or unaware of the visual signal, given the same physical input. We found that perceptual awareness of the visual signal strongly modulated the ventriloquist effect. If participants reported to be unaware of the visual signal, the ventriloquist effect was strongly reduced as compared to trials in which participants reported awareness. However, the visual bias was still significant in case of unaware visual signals.

We concluded that perceptual awareness might modulate the ventriloquist effect via two mechanisms, a cortical and a subcortical pathway: Via a cortical pathway, CFS might strongly decrease the reliability of the visual signals in most trials. Thus, the visual signals were rendered unaware and, according to the MLE model, only induced a weak ventriloquist effect. In trials in which CFS was not successful, the visual signals were more reliable and, therefore, became aware as well as strongly biased the perceived sound location. Alternatively, according to the CI model, invisible flashes might concomitantly lead the participants to infer a low probability of a common cause of the signals. Thereby, the segregation of the visual signal reduces the ventriloquist effect. Because both alternatives rely on cortical pathways, presumably the audiovisual spatial processing hierarchies into IPS, CFS modulates audiovisual integration on a cortical route. On the other hand, the cortical pathway could be entirely blocked if CFS was successful. Thus, only subcortical pathways, for example thalamo-cortical multisensory pathways (Cappe et al., 2009) which cannot be blocked by CFS, mediate a weak visual bias in those trials.

1.6 General discussion and outlook

The results from the PhD projects confirmed many predictions of the CI model, but also call for an extension of the model, and for the first time linked the model to neural processes.

An extended model of causal inference in multisensory perception, linked to the brain

The PhD thesis investigated how human observers perceive and how their cortical hierarchies process multisensory signals (Fig. 1A). In other words, we investigated how physical signals, such as audiovisual spatial signals, map into distinct neuronal representations at different levels of cortical hierarchies. The neuronal multisensory representations in high-level regions like IPS eventually define the subjective perceptual space and, therefore, give rise to perceptual illusion like the ventriloquist effect.

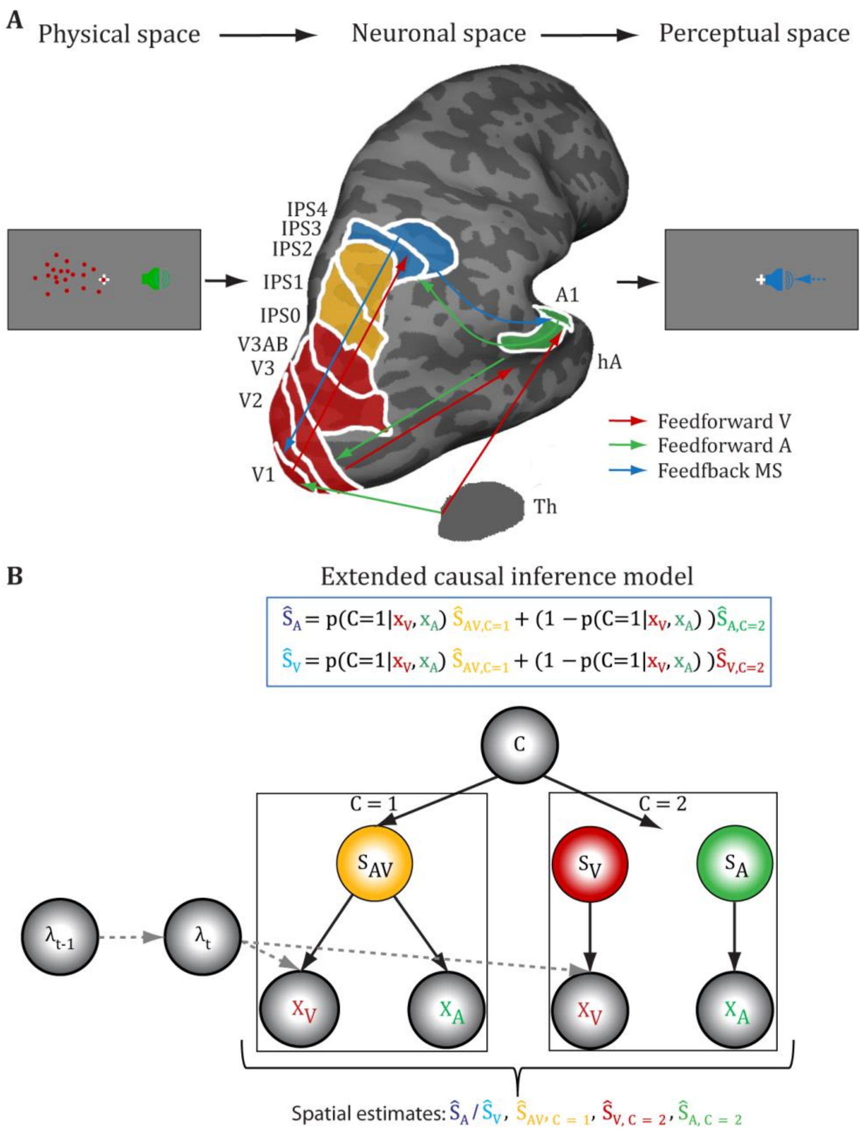

A
Physical space
Neuronal space
Perceptual space
IPS4
IPS3
IPS2
IPS1
IPS0
V3AB
V3
V2
V1
A1
hA
Th
Feedforward V
Feedforward A
Feedfback MS
B
Extended causal inference model
$\hat{S}_A = p(C=1|x_V, x_A)\ \hat{S}_{AV,C=1} + (1 - p(C=1|x_V, x_A))\hat{S}_{A,C=2}$
$\hat{S}_V = p(C=1|x_V, x_A)\ \hat{S}_{AV,C=1} + (1 - p(C=1|x_V, x_A))\hat{S}_{V,C=2}$
C
C = 1
C = 2
S_{AV}
S_V
S_A
λ_{t-1}
λ_t
x_V
x_A
x_V
x_A
Spatial estimates: $\hat{S}_A / \hat{S}_V, \hat{S}_{AV,C=1}, \hat{S}_{V,C=2}, \hat{S}_{A,C=2}$

Figure 1. General framework of audiovisual spatial integration and extended causal inference model. (A) Audiovisual spatial signals are mapped to visual (V) and auditory (A) spatial processing hierachies projecting to intraparietal sulcus (IPS). Along the hierarchy, crossmodal influences are mediated by feedforward cortico-cortical and thalamo-cortical connections as well as multisensory (MS) feedback from higher-level regions. Perceptually, neural processing gives rise to the ventriloquist illusion. V1-3, V3AB = low-level and intermediate visual regions, A1 = primary auditory cortex, hA = higher auditory cortex comprising the planum temporale, Th = thalamus. **(B)** The causal inference model (Kording et al., 2007) computes the posterior auditory and visual spatial estimates ($\hat{S}_A$, $\hat{S}_V$) by weighting the reliability-weighted average ($\hat{S}_{AV,C=1}$) and the task-relevant unisensory spatial estimates ($\hat{S}_{A,C=2}$, $\hat{S}_{V,C=2}$) proportional to the probability of a common ($p(C = 1|x_A, x_V)$) and separate sources ($1 - p(C = 1|x_A, x_V)$). Regions possibly encoding these spatial estimates are color coded in (A) (cf. ch. 3). Further, posterior sensory reliability ($\hat{\lambda}_{,t}$) is learned by updating prior reliablity (λ_{t-1}) with current reliablity information (cf. ch. 6).

Crucially, we employed the Bayesian causal inference model to characterize the signals' mapping from physical to neural to perceptual spaces (Fig. 1B). We found that human observers estimate the signals' causal structure ($\hat{C}$) from signal discrepancies, modulated by sensory reliability (chapter 2). The signals' posterior spatial estimates ($\hat{S}_A$, $\hat{S}_V$) are computed by combining the reliability-weighted estimate ($\hat{S}_{AV, C = 1}$) with the task-relevant unisensory estimate ($\hat{S}_{A, C = 2}$ or $\hat{S}_{V, C = 2}$) according to the probability of common and separate sources, respectively. Thus, if observes explicitly infer a common source, the ventriloquist effect and the variance of the estimate depend on relative sensory reliability, otherwise the signals are mostly segregated (ch. 2).

Importantly, human observers do not exclusively estimate sensory reliability ($\hat{\lambda}_t$) from present signals to weight multisensory signals proportionally: They learn sensory reliability over a time window of several seconds by updating prior estimates of reliability (λ_{t-1}) with new sensory information on reliability (ch. 6).

In the brain, low-level sensory regions (V1-3, A1) mostly represent their preferred unisensory signal estimates, even though we detected crossmodal information from non-preferred sensory channels (ch. 4, 5). The low-level crossmodal information might arise from top-down feedback from IPS, direct connectivity between the sensory regions or even crossmodal thalamo-cortical connections (Fig. 1A). Regions at intermediate stages of the cortical hierarchy (IPS0-2) represent the reliability-weighted estimate (ch. 3). Finally, regions at the top of the hierarchies (IPS3-4) represent the final signal estimates ($\hat{S}_A$, $\hat{S}_V$). Therefore, these regions consider the signals' causal structure (ch. 3) and responds to visual reliability as well as task-relevance especially if signal disparity is large (ch. 4-5). Observers base their behavioral localization responses on the spatial estimates of these regions (ch. 5).

Finally, continuous flash suppression (CFS) disturbs cortical feedforward processing from visual to IPS regions (ch. 7). Thus, the ventriloquist effect is strongly reduced if visual signals are rendered unaware by CFS, presumably due to a concomitant drop of visual reliability or the probability of a common source. A residual, much weaker ventriloquist effect may also be mediated by thalamo-cortical connections.

Naturally, many open questions remain at all levels of analysis. The questions pertain to the causal inference model and its neural basis, but also touch on the relation between multisensory integration and attention, the generalizability and applicability of the findings and methodological issues.

Open questions

The CI model

Even though the CI model explains many multisensory phenomena, several details of the model and the incorporation of further multisensory phenomena remain unclear. Further, the implementation of the CI model in the brain awaits elaboration.

First, it is controversial which decision strategies observers use to balance the signal estimates of the two causal structures by their probability (chapter 2 vs. Wozny et al. (2010)). Various optimal decision strategies exist under assumption of different cost functions (e.g., model averaging minimizes the error of the signal estimate; model selection minimizes the error of the signal and causal estimates). Even the usage of non-optimal decision strategies has been reported (i.e., probability matching) (Wozny et al., 2010). Thus, the use of a specific decision strategy might depend on further cognitive factors like the choice of a cost function. It would be interesting to explore how participants implicitly choose decision strategies and whether the choice can be manipulated by specific instructions. Unfortunately, the predictions of the different decision strategies are highly correlated. Thus, their investigation first needs experimental ideas on how to orthogonalize the decision functions, for example by creating experimental conditions which evoke a common-cause prior of around 0.5.

Second, the current CI model can well model causal inferences based on the signals' spatial discrepancy and reliability (chapter 2). However, observers infer the causal structure of multisensory signals from further structural and cognitive factors like the signals' temporal synchronicity or semantic congruence (Radeau and Bertelson, 1977; Welch and Warren, 1980; Wallace et al., 2004). Hence, it would be interesting to test whether the CI model can jointly account for multiple cues to causal structure which requires extensions of the model. For example, the likelihoods of the sensory estimates (X_A and X_V in Fig. 1B) could be modeled by bivariate Gaussians describing that the signals are differently distributed in space and time under the assumption of common and separate causes (similar to Sato et al. (2007)). Cognitive factors like semantic congruence might be more difficult to model because they are discrete variables (e.g., require a mixture model of Gaussians for the likelihoods).

Third, chapters 3-5 determined where the brain represents the spatial components of the CI model, but it remains unknown which brain regions compute the causal structure's posterior probability ($p(C = 1|x_A, x_V)$) as well as its prior probability. These quantities are crucial to determine the final posterior signal estimates and explicit causal judgments. Frequently, perceptual decisions (e.g., visual motion discrimination) recruit frontal regions which compute decision variables by integrating information from sensory neurons representing the competing alternatives (Heekeren et al., 2008). Thus, it seems plausible that frontal regions are involved in deciding on a causal structure. At any rate, because the prior as well as the posterior probability of the causal structures are inherently correlated with the spatial estimates, special care has to be taken to dissociate these components of the CI model (e.g., by post-hoc trial resampling such that fMRI signals from trials of common and separate-cause judgments are compared based on the same distributions of physical input).

Fourth, the specific mechanism of causal inferences at the level of single neurons and populations of neurons are largely unexplored. According to the theory of probabilistic population codes (PPCs) (Ma et al., 2006), summation over unisensory PPCs implements reliability-weighted integration for the case of common sources. Chapters 3-5 suggest that further downstream regions (e.g., anterior IPS) combine such a population response profile with the task-relevant unisensory profile biased by information on the causal structure. Thus, future computational and neurophysiological studies could investigate how causal inference is computationally feasible and actually implemented in neural populations.

Neural representations of reliability

Chapter 6 revealed that the brain estimates sensory reliability not only from current signals as assumed by the MLE (Ernst and Banks, 2002) and PPC (Ma et al., 2006) theories. By contrast, human observers consider the recent history of sensory reliability as a Bayesian prior. Therefore, it is unclear how this prior reliability is encoded by the brain and integrated with current estimates of reliability. PPC theory could be extended by assuming that prior reliability modulates the populations' response gain representing current reliability. Alternatively, the brain could use a sampling-based representation in which the neurons' activity encodes samples of a signal (Fiser et al., 2010). Thus, reliability is computed via the variability of neural responses over time and, therefore, naturally history-dependent. In any case, chapter 6 provided new empirical constraints to models of neuronal population implementing Bayesian inference.

An important question for future studies is which brain regions encode prior, current and posterior reliability. In an fMRI study, these reliability estimates could be derived from fitting the Bayesian learner of chapter 6 to behavioral data. To determine brain regions encoding reliability estimates, these estimates could be regressed against (i.e., mass-univariate analysis) or decoded from BOLD responses (i.e., multivariate analysis). Crucially, this approach would have to take the collinearity of the reliability estimates into account.

The relation of multisensory integration and attention

Currently, it is controversial whether top-down factors like attention and task relevance influence multisensory integration or whether multisensory integration is immune to such factors (Talsma et al., 2010). Chapters 3-5 support the notion of an interplay between multisensory integration and attention because they demonstrate that the task-relevance of a signal strongly influences integration at the perceptual level and in higher association cortex. The CI model implicitly also assumes an interplay because it is the *task-relevant* unisensory signal estimate which is combined with the reliability-weighted estimate when integrating over the potential causal structures. Moreover, chapter 4 also demonstrated that the spatial estimates in anterior IPS are behaviorally relevant (i.e., used for overt localization responses). Thus, the chapter suggested that IPS builds a multisensory priority map in which a relevant location in space is jointly defined across modalities by bottom-up (i.e., sensory reliability) and top-down (i.e., task-relevance) factors. This notion extends the model of unisensory visual priority maps (Itti and Koch, 2001; Bisley and Goldberg, 2010; Ptak, 2012) which determine relevant locations in visual space by integrating multiple visual 'feature maps' (e.g., for color or edge orientation) biased by top-down goals (e.g., intentionally focusing on a certain color).

However, chapter 4 only showed that IPS represents the location which is relevant for overt localization responses, but the studies did not investigate whether these locations are indeed selected by covert attention. Therefore, it might be unclear whether multisensory priority maps represent the location only for motor intentions (Andersen et al., 1997) and/or attentional selection (Colby and Goldberg, 1999). An experiment using the orthogonal spatial cueing paradigm (Spence and Driver, 1994) could investigate whether multisensory cues create specific cueing effects (i.e., smaller reaction times to classify a spatial target after a valid vs. invalid spatial cue). If observers build a multisensory representation of the cues according to the CI model, the cueing effects should depend on the reliability and task-relevance (for a secondary spatial localization task) of the cues. In fact, the CI model could describe the multisensory cue representation. Moreover, an extension of the model could describe the attentional selection of the target which depends on the distance of the multisensory cue representation to the target.

Using multivariate decoding approaches to determine the neural basis of the ventriloquist after-effect.

The PhD projects exclusively investigated the 'online' integration of multisensory signals which are processed by 'accurate' sensors (i.e., sensors which represent a physical quantity unbiased on average). However, if for example sensory organs grow or are damaged, the senses may lose their accuracy and provide biased information. In this case, another modality can restore the accuracy of the biased percept by recalibrating it. Recalibration leads to 'offline' aftereffects: For example, after audiovisual spatial signals have been presented with a constant discrepancy for an extended time, unisensory auditory signals are perceived shifted in direction of the previously presented visual signals (Radeau and

Bertelson, 1974). This ‘ventriloquism aftereffect’ reveals that the auditory percepts were recalibrated by the visual signals. However, the neural basis of the ventriloquist aftereffect remains unknown.

Multivariate fMRI decoding approaches using regression models (cf. chapter 3-4) are well suited for such an investigation. For example, using fMRI activation patterns from a pre-training unimodal auditory session, one could train a regression model to learn the mapping from the patterns to the presented auditory signal locations. Using patterns from a post-training auditory session, the trained regression model might decode auditory locations which are shifted in direction of the visual signals previously presented during training. Thus, the approach could pinpoint at which level of the auditory processing hierarchy the ventriloquist aftereffect emerges.

More generally, multivariate fMRI decoding approaches as used in chapter 3-5 are a powerful tool to investigate the mapping between physical, neuronal and perceptual spaces for many kinds of signals (cf. Fig. 1A). This is because the approaches map a high-dimensional space of fMRI activation patterns to a unidimensional variable encoding the neurally processed physical signal within a region. Thus, this ‘neural’ variable can be analyzed as a function of the experimental conditions in the same way as the perceptual variable (e.g., a perceived location), and related to it. Crucially, this approach tracks how neural spaces at different levels of processing hierarchies transform the physical input. By comparing these neural spaces with the perceptual space, one can pinpoint which regions likely host a close correlate of the subjective perception. For example, this approach could be applied to compare psychometric and ‘neurometric’ functions in unisensory signal integration (Ban et al., 2012), perceptual adaptation (Tootell et al., 1995) or representation of ‘face space’ (Loffler et al., 2005).

Generalizability of the findings

All projects were conducted using highly artificial audiovisual spatial signals. It is speculative to infer that similar results would be obtained with other combinations of multisensory signals and with more natural multisensory stimuli. However, it is remarkable that the MLE as well as the CI model were successfully applied to many domains of signal combination. In the multisensory and unisensory domains, the models were for example applied to visual-haptic integration (Ernst and Banks, 2002; Hospedales and Vijayakumar, 2009) and estimation of visual depth from visual disparity and texture (Knill and Saunders, 2003; Knill, 2007). The models were even applied in sensorimotor control (Kording and Wolpert, 2004; Wei and Kording, 2009). Therefore, the results of the current studies likely generalize to diverse psychophysical functions and neural processes which deal with noisy sensory signals, but this claim obviously awaits future empirical support.

The question of generalizability to natural multisensory stimuli is more difficult because model-based studies tend to choose well-controlled artificial stimuli which can be easily linked to model parameters. It is striking that older studies on multisensory integration which used comparably 'model-free' analyses often employed more ecologically valid stimuli (e.g., a whistling sound and a steaming kettle to investigate audiovisual spatial integration (Jackson, 1953)). Thus, the ecological validity of the CI model should be definitely explored in more natural conditions, for example when multiple visual and auditory signals complicate causal inference even further.

Applicability of the CI model

The results of the PhD projects as well as the CI model in general could be applicable to technical problems as well as for diagnosis and treatment of psychiatric and neurological disorders. As an example of a technical problem, the tracking of objects in a multisensory scene (e.g., several persons at a telephone conference) has been tackled by inferring the causal structure of such multisensory data (Hospedales and Vijayakumar, 2008). Further, clinical studies found altered multisensory integration in psychiatric disorders like schizophrenia (de Gelder et al., 2003) or autism (Foss-Feig et al., 2010; Brandwein et al., 2013) and in neurological disorders like hemianopia (Leo et al., 2008). These studies mostly used simple multisensory paradigms and for example compared the intersensory bias or multisensory redundancy effects between patients and healthy controls. By contrast, the CI model gives a detailed description of multisensory processes. The model allows more fine-grained analyses of impaired multisensory processes in disorders, for example by comparing CI model parameters between patients and controls. Because the current PhD projects also pinpoint these processes to brain regions, future clinical studies could also test whether patients show altered multisensory neural processing in these regions. Thus, models of disorders could be extended to include multisensory 'symptoms' which could be helpful to define further diagnostic criteria and targets for interventions.

1.7 Conclusions

Humans effortlessly build complex multisensory representation of their environment. The results of the PhD thesis demonstrate that such multisensory representations arise from cortical hierarchies which employ Bayesian principles like causal inference and reliability-weighted integration to combine or to segregate multisensory signals according to their causal structure. From a Bayesian perspective (Yuille and Buelthoff, 1996; Knill and Pouget, 2004; Yuille and Kersten, 2006), the results demonstrate that the brain infers physical quantities from their noisy, uncertain signals to build an ‘optimal’, best-informed multisensory representation of the environment. Thus, the PhD thesis adds to the growing body of evidence demonstrating Bayesian principles in perception and the brain (notwithstanding current critical arguments against ‘Bayesianism’ (Jones and Love, 2011; Bowers and Davis, 2012)).

For neural theories of multisensory integration (Ghazanfar and Schroeder, 2006; Driver and Noesselt, 2008), the most important finding is that not a single multisensory region, but cortical hierarchies jointly implement multisensory causal inference processes. Thus, the PhD thesis suggests that multisensory integration in the brain can only be fully understood if specific multisensory processes are investigated simultaneously in cortical hierarchies or even larger networks comprising subcortical structures.

Hence, the most urgent open question might be how unisensory and multisensory populations of neurons implement the causal inference model. To tackle this question, it requires new computational theories of neural mechanisms of causal inference and neurophysiological data simultaneously recorded from remote populations of neurons.

1.8 References

Alais D, Burr D (2004) The ventriloquist effect results from near-optimal bimodal integration. Curr Biol 14:257-262.

Alsius A, Navarra J, Soto-Faraco S (2007) Attention to touch weakens audiovisual speech integration. Exp Brain Res 183:399-404.

Alsius A, Navarra J, Campbell R, Soto-Faraco S (2005) Audiovisual integration of speech falters under high attention demands. Curr Biol 15:839-843.

Andersen RA, Snyder LH, Bradley DC, Xing J (1997) Multimodal representation of space in the posterior parietal cortex and its use in planning movements. Annu Rev Neurosci 20:303-330.

Avillac M, Deneve S, Olivier E, Pouget A, Duhamel JR (2005) Reference frames for representing visual and tactile locations in parietal cortex. Nat Neurosci 8:941-949.

Ban H, Preston TJ, Meeson A, Welchman AE (2012) The integration of motion and disparity cues to depth in dorsal visual cortex. Nat Neurosci 15:636-643.

Battaglia PW, Jacobs RA, Aslin RN (2003) Bayesian integration of visual and auditory signals for spatial localization. J Opt Soc Am A Opt Image Sci Vis 20:1391-1397.

Beauchamp MS, Argall BD, Bodurka J, Duyn JH, Martin A (2004) Unraveling multisensory integration: patchy organization within human STS multisensory cortex. Nat Neurosci 7:1190-1192.

Bertelson P, Radeau M (1981) Cross-modal bias and perceptual fusion with auditory-visual spatial discordance. Attention, Perception, & Psychophysics 29:578-584.

Bertelson P, Vroomen J, de Gelder B, Driver J (2000a) The ventriloquist effect does not depend on the direction of deliberate visual attention. Percept Psychophys 62:321-332.

Bertelson P, Pavani F, Ladavas E, Vroomen J, de Gelder B (2000b) Ventriloquism in patients with unilateral visual neglect. Neuropsychologia 38:1634-1642.

Bisley JW, Goldberg ME (2010) Attention, intention, and priority in the parietal lobe. Annu Rev Neurosci 33:1-21.

Bizley JK, Nodal FR, Bajo VM, Nelken I, King AJ (2007) Physiological and anatomical evidence for multisensory interactions in auditory cortex. Cereb Cortex 17:2172-2189.

Botvinick M, Cohen J (1998) Rubber hands 'feel' touch that eyes see. Nature 391:756.

Bowers JS, Davis CJ (2012) Bayesian just-so stories in psychology and neuroscience. Psychol Bull 138:389-414.

Brandwein AB, Foxe JJ, Butler JS, Russo NN, Altschuler TS, Gomes H, Molholm S (2013) The development of multisensory integration in high-functioning autism: high-density electrical mapping and psychophysical measures reveal impairments in the processing of audiovisual inputs. Cereb Cortex 23:1329-1341.

Bremmer F, Schlack A, Shah NJ, Zafiris O, Kubischik M, Hoffmann K, Zilles K, Fink GR (2001) Polymodal motion processing in posterior parietal and premotor cortex: a human fMRI study strongly implies equivalencies between humans and monkeys. Neuron 29:287-296.

Bresciani JP, Dammeier F, Ernst MO (2006) Vision and touch are automatically integrated for the perception of sequences of events. J Vis 6:554-564.

Bruce C, Desimone R, Gross CG (1981) Visual properties of neurons in a polysensory area in superior temporal sulcus of the macaque. J Neurophysiol 46:369-384.

Cappe C, Morel A, Barone P, Rouiller EM (2009) The thalamocortical projection systems in primate: an anatomical support for multisensory and sensorimotor interplay. Cereb Cortex 19:2025-2037.

Colby CL, Goldberg ME (1999) Space and attention in parietal cortex. Annu Rev Neurosci 22:319-349.

Dahl CD, Logothetis NK, Kayser C (2009) Spatial organization of multisensory responses in temporal association cortex. J Neurosci 29:11924-11932.

de Gelder B, Vroomen J, Annen L, Masthof E, Hodiamont P (2003) Audio-visual integration in schizophrenia. Schizophr Res 59:211-218.

Diederich A, Colonius H (2004) Bimodal and trimodal multisensory enhancement: effects of stimulus onset and intensity on reaction time. Percept Psychophys 66:1388-1404.

Donohue SE, Roberts KC, Grent-'t-Jong T, Woldorff MG (2011) The cross-modal spread of attention reveals differential constraints for the temporal and spatial linking of visual and auditory stimulus events. J Neurosci 31:7982-7990.

Driver J (1996) Enhancement of selective listening by illusory mislocation of speech sounds due to lip-reading. Nature 381:66-68.

Driver J, Noesselt T (2008) Multisensory interplay reveals crossmodal influences on 'sensory-specific' brain regions, neural responses, and judgments. Neuron 57:11-23.

Duhamel JR, Colby CL, Goldberg ME (1998) Ventral intraparietal area of the macaque: congruent visual and somatic response properties. J Neurophysiol 79:126-136.

Ernst MO, Banks MS (2002) Humans integrate visual and haptic information in a statistically optimal fashion. Nature 415:429-433.

Ernst MO, Bulthoff HH (2004) Merging the senses into a robust percept. Trends Cogn Sci 8:162-169.

Fairhall SL, Macaluso E (2009) Spatial attention can modulate audiovisual integration at multiple cortical and subcortical sites. Eur J Neurosci 29:1247-1257.

Faisal AA, Selen LP, Wolpert DM (2008) Noise in the nervous system. Nat Rev Neurosci 9:292-303.

Falchier A, Clavagnier S, Barone P, Kennedy H (2002) Anatomical evidence of multimodal integration in primate striate cortex. J Neurosci 22:5749-5759.

Fetsch CR, Deangelis GC, Angelaki DE (2013) Bridging the gap between theories of sensory cue integration and the physiology of multisensory neurons. Nat Rev Neurosci 14:429-442.

Fetsch CR, Pouget A, DeAngelis GC, Angelaki DE (2012) Neural correlates of reliability-based cue weighting during multisensory integration. Nat Neurosci 15:146-154.

Fiser J, Berkes P, Orban G, Lengyel M (2010) Statistically optimal perception and learning: from behavior to neural representations. Trends Cogn Sci 14:119-130.

Foss-Feig JH, Kwakye LD, Cascio CJ, Burnette CP, Kadivar H, Stone WL, Wallace MT (2010) An extended multisensory temporal binding window in autism spectrum disorders. Exp Brain Res 203:381-389.

Foxe JJ, Morocz IA, Murray MM, Higgins BA, Javitt DC, Schroeder CE (2000) Multisensory auditory-somatosensory interactions in early cortical processing revealed by high-density electrical mapping. Brain Res Cogn Brain Res 10:77-83.

Gepshtein S, Banks MS (2003) Viewing geometry determines how vision and haptics combine in size perception. Curr Biol 13:483-488.

Gepshtein S, Burge J, Ernst MO, Banks MS (2005) The combination of vision and touch depends on spatial proximity. J Vis 5:1013-1023.

Ghazanfar AA, Schroeder CE (2006) Is neocortex essentially multisensory? Trends Cogn Sci 10:278-285.

Haxby JV, Gobbini MI, Furey ML, Ishai A, Schouten JL, Pietrini P (2001) Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science 293:2425-2430.

Heekeren HR, Marrett S, Ungerleider LG (2008) The neural systems that mediate human perceptual decision making. Nat Rev Neurosci 9:467-479.

Helbig HB, Ernst MO (2008) Visual-haptic cue weighting is independent of modality-specific attention. J Vis 8:21 21-16.

Helbig HB, Ernst MO, Ricciardi E, Pietrini P, Thielscher A, Mayer KM, Schultz J, Noppeney U (2012) The neural mechanisms of reliability weighted integration of shape information from vision and touch. Neuroimage 60:1063-1072.

Hershenson M (1962) Reaction time as a measure of intersensory facilitation. Journal of Experimental Psychology 63:289.

Hillis JM, Watt SJ, Landy MS, Banks MS (2004) Slant from texture and disparity cues: optimal cue combination. J Vis 4:967-992.

Hospedales T, Vijayakumar S (2009) Multisensory oddity detection as bayesian inference. PLoS One 4:e4205.

Hospedales TM, Vijayakumar S (2008) Structure inference for Bayesian multisensory scene understanding. IEEE Trans Pattern Anal Mach Intell 30:2140-2157.

Itti L, Koch C (2001) Computational modelling of visual attention. Nat Rev Neurosci 2:194-203.

Jack CE, Thurlow WR (1973) Effects of degree of visual association and angle of displacement on the "ventriloquism" effect. Percept Mot Skills 37:967-979.

Jackson C (1953) Visual factors in auditory localization. Quarterly Journal of Experimental Psychology 5:52-65.

Jacobs RA (1999) Optimal integration of texture and motion cues to depth. Vision Res 39:3621-3629.

Jones M, Love BC (2011) Bayesian Fundamentalism or Enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. Behav Brain Sci 34:169-188; disuccsion 188-231.

Jousmaki V, Hari R (1998) Parchment-skin illusion: sound-biased touch. Curr Biol 8:R190.

Kalman RE (1960) A new approach to linear filtering and prediction problems. Journal of basic Engineering 82:35-45.

Kay KN, Gallant JL (2009) I can see what you see. Nat Neurosci 12:245.

Kayser C, Petkov CI, Augath M, Logothetis NK (2007) Functional imaging reveals visual modulation of specific fields in auditory cortex. J Neurosci 27:1824-1835.

Knill DC (2007) Robust cue integration: a Bayesian model and evidence from cue-conflict studies with stereoscopic and figure cues to slant. J Vis 7:5 1-24.

Knill DC, Saunders JA (2003) Do humans optimally integrate stereo and texture information for judgments of surface slant? Vision Res 43:2539-2558.

Knill DC, Pouget A (2004) The Bayesian brain: the role of uncertainty in neural coding and computation. Trends Neurosci 27:712-719.

Kording KP, Wolpert DM (2004) Bayesian integration in sensorimotor learning. Nature 427:244-247.

Kording KP, Beierholm U, Ma WJ, Quartz S, Tenenbaum JB, Shams L (2007) Causal inference in multisensory perception. PLoS One 2:e943.

Lakatos P, Chen CM, O'Connell MN, Mills A, Schroeder CE (2007) Neuronal oscillations and multisensory interaction in primary auditory cortex. Neuron 53:279-292.

Lenggenhager B, Tadi T, Metzinger T, Blanke O (2007) Video ergo sum: manipulating bodily self-consciousness. Science 317:1096-1099.

Leo F, Bolognini N, Passamonti C, Stein BE, Ladavas E (2008) Cross-modal localization in hemianopia: new insights on multisensory integration. Brain 131:855-865.

Lewald J, Guski R (2003) Cross-modal perceptual integration of spatially and temporally disparate auditory and visual stimuli. Brain Res Cogn Brain Res 16:468-478.

Lewis R, Noppeney U (2010) Audiovisual synchrony improves motion discrimination via enhanced connectivity between early visual and auditory areas. J Neurosci 30:12329-12339.

Loffler G, Yourganov G, Wilkinson F, Wilson HR (2005) fMRI evidence for the neural representation of faces. Nat Neurosci 8:1386-1390.

Ma WJ, Beck JM, Latham PE, Pouget A (2006) Bayesian inference with probabilistic population codes. Nat Neurosci 9:1432-1438.

Macaluso E, Driver J (2005) Multisensory spatial interactions: a window onto functional integration in the human brain. Trends Neurosci 28:264-271.

McDonald JJ, Teder-Salejarvi WA, Hillyard SA (2000) Involuntary orienting to sound improves visual perception. Nature 407:906-908.

McGurk H, MacDonald J (1976) Hearing lips and seeing voices. Nature 264:746-748.

Meredith MA, Stein BE (1983) Interactions among converging sensory inputs in the superior colliculus. Science 221:389-391.

Meredith MA, Stein BE (1986a) Visual, auditory, and somatosensory convergence on cells in superior colliculus results in multisensory integration. J Neurophysiol 56:640-662.

Meredith MA, Stein BE (1986b) Spatial factors determine the activity of multisensory neurons in cat superior colliculus. Brain Res 365:350-354.

Meredith MA, Nemitz JW, Stein BE (1987) Determinants of multisensory integration in superior colliculus neurons. I. Temporal factors. J Neurosci 7:3215-3229.

Miller J (1982) Divided attention: evidence for coactivation with redundant signals. Cogn Psychol 14:247-279.

Mishkin M, Ungerleider LG, Macko KA (1983) Object vision and spatial vision: two cortical pathways. Trends in neurosciences 6:414-417.

Molholm S, Ritter W, Murray MM, Javitt DC, Schroeder CE, Foxe JJ (2002) Multisensory auditory-visual interactions during early sensory processing in humans: a high-density electrical mapping study. Brain Res Cogn Brain Res 14:115-128.

Morein-Zamir S, Soto-Faraco S, Kingstone A (2003) Auditory capture of vision: examining temporal ventriloquism. Brain Res Cogn Brain Res 17:154-163.

Morgan ML, Deangelis GC, Angelaki DE (2008) Multisensory integration in macaque visual cortex depends on cue reliability. Neuron 59:662-673.

Parise CV, Spence C, Ernst MO (2012) When correlation implies causation in multisensory integration. Curr Biol 22:46-49.

Pick HL, Warren DH, Hay JC (1969) Sensory conflict in judgments of spatial direction. Perception & Psychophysics 6:203-205.

Pouget A, Deneve S, Duhamel JR (2002) A computational perspective on the neural basis of multisensory spatial representations. Nat Rev Neurosci 3:741-747.

Ptak R (2012) The frontoparietal attention network of the human brain: action, saliency, and a priority map of the environment. Neuroscientist 18:502-515.

Radeau M (1985) Signal intensity, task context, and auditory-visual interactions. Perception 14:571-577.

Radeau M, Bertelson P (1974) The after-effects of ventriloquism. Q J Exp Psychol 26:63-71.

Radeau M, Bertelson P (1977) Adaptation to auditory-visual discordance and ventriloquism in semirealistic situations. Perception & Psychophysics 22:137-146.

Recanzone GH, Sutter ML (2008) The biological basis of audition. Annu Rev Psychol 59:119-142.

Roach NW, Heron J, McGraw PV (2006) Resolving multisensory conflict: a strategy for balancing the costs and benefits of audio-visual integration. Proc Biol Sci 273:2159-2168.

Rock I, Victor J (1964) Vision and Touch: An Experimentally Created Conflict between the Two Senses. Science 143:594-596.

Sadaghiani S, Maier JX, Noppeney U (2009) Natural, metaphoric, and linguistic auditory direction signals have distinct influences on visual motion processing. J Neurosci 29:6490-6499.

Sato Y, Toyoizumi T, Aihara K (2007) Bayesian inference explains perception of unity and ventriloquism aftereffect: identification of common sources of audiovisual stimuli. Neural Comput 19:3335-3355.

Serences JT, Saproo S (2012) Computational advances towards linking BOLD and behavior. Neuropsychologia 50:435-446.

Shams L, Beierholm UR (2010) Causal inference in perception. Trends Cogn Sci 14:425-432.

Shams L, Kamitani Y, Shimojo S (2000) Illusions. What you see is what you hear. Nature 408:788.

Shipley T (1964) Auditory Flutter-Driving of Visual Flicker. Science 145:1328-1330.

Spence CJ, Driver J (1994) Covert spatial orienting in audition: Exogenous and endogenous mechanisms. Journal of Experimental Psychology: Human Perception and Performance 20:555.

Stein BE, Meredith MA (1993) The merging of the senses. Cambridge, MA: The MIT Press.

Stein BE, Stanford TR (2008) Multisensory integration: current issues from the perspective of the single neuron. Nat Rev Neurosci 9:255-266.

Talsma D, Senkowski D, Soto-Faraco S, Woldorff MG (2010) The multifaceted interplay between attention and multisensory integration. Trends Cogn Sci 14:400-410.

Thomas GJ (1941) Experimental study of the influence of vision on sound localization. Journal of Experimental Psychology 28:163.

Tian B, Reser D, Durham A, Kustov A, Rauschecker JP (2001) Functional specialization in rhesus monkey auditory cortex. Science 292:290-293.

Tootell RB, Reppas JB, Dale AM, Look RB, Sereno MI, Malach R, Brady TJ, Rosen BR (1995) Visual motion aftereffect in human cortical area MT revealed by functional magnetic resonance imaging. Nature 375:139-141.

Tsuchiya N, Koch C (2005) Continuous flash suppression reduces negative afterimages. Nat Neurosci 8:1096-1101.

Van der Burg E, Olivers CN, Bronkhorst AW, Theeuwes J (2008) Pip and pop: nonspatial auditory signals improve spatial visual search. J Exp Psychol Hum Percept Perform 34:1053-1065.

Vroomen J, Bertelson P, de Gelder B (2001) The ventriloquist effect does not depend on the direction of automatic visual attention. Percept Psychophys 63:651-659.

Wallace MT, Roberson GE, Hairston WD, Stein BE, Vaughan JW, Schirillo JA (2004) Unifying multisensory signals across time and space. Exp Brain Res 158:252-258.

Wandell BA, Dumoulin SO, Brewer AA (2007) Visual field maps in human cortex. Neuron 56:366-383.

Warren DH (1979) Spatial localization under conflict conditions: is there a single explanation? Perception 8:323-337.

Warren DH, Cleaves WT (1971) Visual-proprioceptive interaction under large amounts of conflict. J Exp Psychol 90:206-214.

Wei K, Kording K (2009) Relevance of error: what drives motor adaptation? J Neurophysiol 101:655-664.

Welch RB, Warren DH (1980) Immediate perceptual response to intersensory discrepancy. Psychol Bull 88:638-667.

Welch RB, DuttonHurt LD, Warren DH (1986) Contributions of audition and vision to temporal rate perception. Percept Psychophys 39:294-300.

Werner S, Noppeney U (2010) Superadditive responses in superior temporal sulcus predict audiovisual benefits in object categorization. Cereb Cortex 20:1829-1842.

Wozny DR, Beierholm UR, Shams L (2008) Human trimodal perception follows optimal statistical inference. J Vis 8:24 21-11.

Wozny DR, Beierholm UR, Shams L (2010) Probability matching as a computational strategy used in perception. PLoS Comput Biol 6.

Yuille A, Kersten D (2006) Vision as Bayesian inference: analysis by synthesis? Trends Cogn Sci 10:301-308.

Yuille AL, Buelthoff HH (1996) Bayesian decision theory and psychophysics. New York: Cambridge University Press.

1.9 Declaration of contributions

The PhD thesis comprises of six manuscripts which are submitted or will be submitted soon for publication. In the following, the contributions of the candidate and the co-authors are detailed:

1. Rohe T., Noppeney U. (2014) Sensory reliability shapes causal inference via two mechanisms (under review): T.R. collected the data, programmed the experiment and wrote analysis scripts. T.R. and U.N. conceived the experiment and data analysis and wrote the manuscript.
2. Rohe T., Noppeney U. (2014) Cortical hierarchies perform Bayesian causal inference for multisensory perception (under review): T.R. collected the data, programmed the experiment and wrote analysis scripts. T.R. and U.N. conceived the experiment and data analysis and wrote the manuscript.
3. Rohe T., Noppeney U. (2014) To integrate, or not to integrate: Causal inference in primary sensory and association cortices during multisensory perception: T.R. collected the data, programmed the experiment and wrote analysis scripts. T.R. and U.N. conceived the experiment and data analysis and wrote the manuscript.
4. Rohe T., Noppeney U. (2014) Suboptimal reliability-weighted integration of audiovisual spatial signals in parietal cortex: T.R. collected the data, programmed the experiment and wrote analysis scripts. T.R. and U.N. conceived the experiment and data analysis and wrote the manuscript.
5. Rohe T.*, Beierholm U.*, Stegle O., Noppeney U. (2014) Bayesian learning of sensory reliability in multisensory perception: T.R. collected the data and programmed the experiment. U.N., T.R. and O.S. conceived the experiment. T.R., U.N. and U.B. analyzed the data and wrote the manuscript.
6. Giani, A. S.*, Rohe T.*, Mátè Aller*, Conrad V., Watanabe M., Noppeney U. (2014) The invisible ventriloquist (under review): A.S.G., V.C., M.W., U.N. designed the study and stimuli. A.S.G. and V.C. acquired the data. A.S.G., T.R., M.A. and U.N. analyzed the data. All authors wrote the manuscript.

* shared first authorship

Parts of this work were also presented at the following conferences:

1. Rohe T., Noppeney U. (2011) The ventriloquist effect depends on audiovisual spatial discrepancy and visual reliability, 12th Conference of Junior Neuroscientists of Tübingen (NeNA), Heiligkreuztal, Germany.

2. Rohe T., Noppeney U. (2012) Intraparietal sulcus represents audiovisual space, Bernstein Conference 2012, München, Germany, Frontiers in Computational Neuroscience, Conference Abstract: Bernstein Conference 2012 192-193.
3. Rohe T., Noppeney U. (2012) Neural audiovisual representations of space in sensory and higher multisensory cortices, 42nd Annual Meeting of the Society for Neuroscience (Neuroscience 2012), New Orleans, LA, USA
4. Rohe T., Noppeney U. (2013) Causal inference conditions reliability-weighted integration of audiovisual spatial signals, Bernstein Conference 2013, Tübingen, Germany.
5. Rohe T., Noppeney U. (2013) Intraparietal sulcus forms multisensory spatial priority maps, 43rd Annual Meeting of the Society for Neuroscience (Neuroscience 2013), San Diego, CA, USA.
6. Rohe T., Noppeney U. (2014): A cortical hierarchy performs Bayesian Causal Inference for multisensory perception, 20th Annual Meeting of the Organization for Human Brain Mapping (OHBM 2014), Hamburg, Germany.

2 Sensory reliability shapes causal inference via two mechanisms

2.1 Abstract

To obtain a coherent percept of the environment, the brain should integrate sensory signals from common and segregate those from independent sources. Recent research has demonstrated that humans integrate audiovisual information during spatial localization consistent with Bayesian causal inference. However, the decision strategies that human observers employ for implicit and explicit causal inference remain unclear. Further, despite the key role of sensory reliability in multisensory integration, Bayesian causal inference has never been evaluated across a wide range of sensory reliabilities. This psychophysics study presented participants with spatially congruent and discrepant audiovisual signals at four levels of visual reliability. Participants localized the auditory signals (implicit causal inference) and judged whether auditory and visual signals come from common or independents sources (explicit causal inference). Our results demonstrate that humans employ model averaging as a decision strategy for implicit causal inference; they report an auditory spatial estimate that averages the spatial estimates under the two causal structures weighted by their posterior probabilities. Likewise, they explicitly infer a common source during the common-source judgment when the posterior probability for a common source exceeds a fixed threshold of 0.5. Critically, sensory reliability shapes multisensory integration in Bayesian Casual inference via two distinct mechanisms: First, higher sensory reliability sensitizes humans to spatial disparity and thereby sharpens their multisensory integration window. Second, sensory reliability determines the relative signal weights in multisensory integration under the assumption of a common source. In conclusion, our results demonstrate that Bayesian causal inference is fundamental for integrating signals of variable reliabilities.

2.2 Introduction

Imagine you are engaged in a conversation at a busy party. You will understand your conversational partner more clearly when you integrate the acoustic speech with his facial articulatory movements. By contrast, speech comprehension will deteriorate if you erroneously integrate his facial movements with another person’s acoustic speech signal. Thus, audiovisual integration requires the brain to infer whether signals come from common or independent sources. This challenge cannot be addressed by traditional forced-fusion models that forces signals to be integrated in a mandatory fashion (Ernst and Banks, 2002) but requires Bayesian causal inference that explicitly models the potential causal structures that could have generated the sensory signals (Kording et al., 2007; Shams and Beierholm, 2010). In the case of a common source, the sensory signals are integrated weighted by their reliability into the most reliable unbiased estimate. In the case of separate sources, signals are processed independently. Importantly, the brain does not know the underlying causal structure, but needs to infer it from the sensory signals based on spatial, temporal and structural correspondences (Slutsky and Recanzone, 2001; Lewald and Guski, 2003; Wallace et al., 2004). A final estimate of a physical property is obtained by combining the estimates under the various causal structures using decisional strategies such as model averaging, model selection or probability matching (Wozny et al., 2010).

Previous modelling efforts have demonstrated that humans integrate information for spatial localization consistent with Bayesian causal inference. For small spatial discrepancies the perceived location of an auditory event shifts towards the location of a temporally correlated but spatially displaced visual event and vice versa depending on the relative auditory and visual reliabilities (Alais and Burr, 2004). Yet, for large spatial discrepancies, when it is unlikely that audiovisual signals arise from a common source, these crossmodal biases are greatly attenuated (Wallace et al., 2004). Moreover, when participants indicated that the audiovisual signals come from independent sources, the perceived auditory location shifted less towards or even away from the true visual location (Wallace et al., 2004; Kording et al., 2007).

However, so far Bayesian causal inference models have been applied to psychophysics data that included only one or two reliability levels (Beierholm et al., 2009). Given the key role of reliability in multisensory integration, it is critical to demonstrate that Bayesian causal inference predicts observers' response profile when sensory signals vary in their reliability over a wide range. Furthermore, it is unclear how participants perform causal inference decisions implicitly during spatial localization and explicitly during common-source judgments. For audiovisual spatial localization, one recent study has suggested that humans do not perform model averaging as previously assumed, but employ a suboptimal strategy of probability matching (Wozny et al., 2010). In other words, they report the spatial estimate of one particular causal structure sampled in proportion to the posterior probability of this causal structure.

Yet, it is unclear whether a similar decision strategy is employed, when causal inference decisions are invoked explicitly in common-source judgments. As implicit and explicit causal inference tasks serve different goals, they may be governed by different utility functions associated with different decision strategies. It is conceivable that implicit and explicit causal inference access the same posterior common-source probability, yet use it with different decision strategies.

To address these questions, we presented participants with spatially congruent and discrepant audiovisual signals at four visual reliability levels in a spatial ventriloquist paradigm. On each trial, participants located the auditory signal and judged whether the audiovisual signals emanated from a common source. We then fitted the Bayesian causal inference model commonly to spatial localization and common-source judgments under various decision strategies.

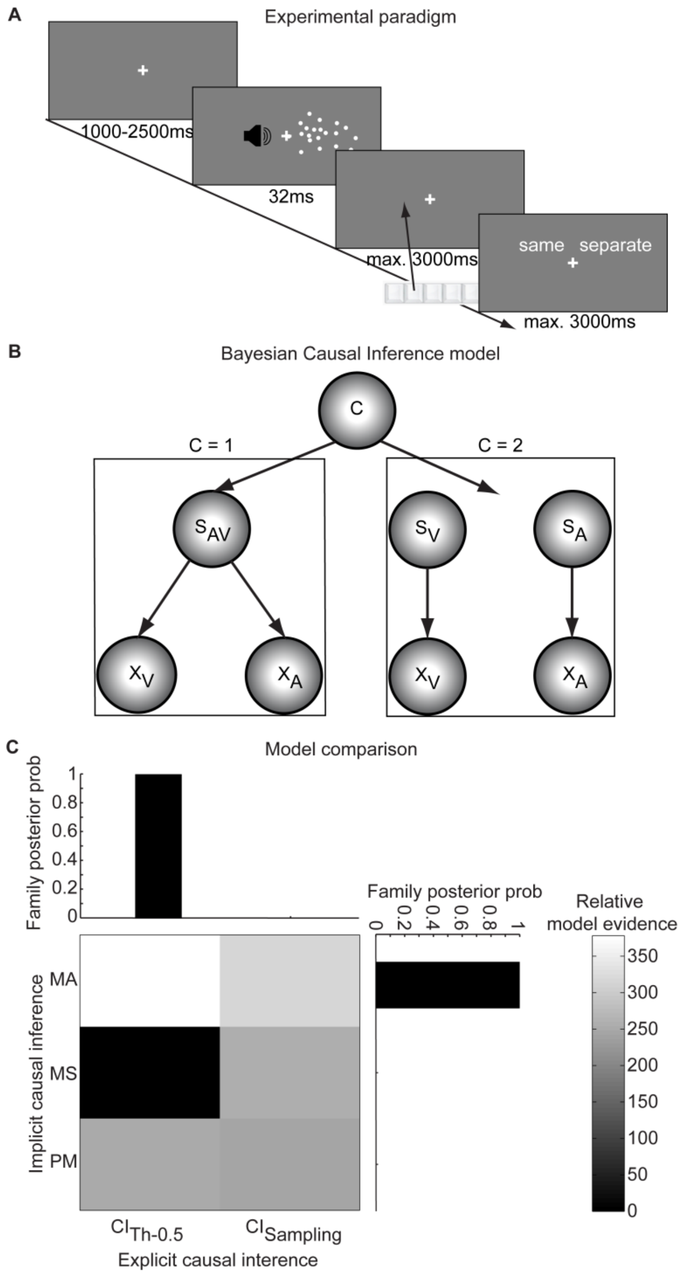


Figure 2.1. Experimental design, Bayesian causal inference model and results of the model comparison. **(A)** Stimuli and time course of an experimental trial in the ventriloquist paradigm. After a variable fixation interval, participants were presented with synchronous, spatially congruent or discrepant visual and auditory signals along the horizontal meridian. Using five response buttons, participants localized the auditory signal and decided whether the visual and auditory signals were generated by common ('same') or independent ('separate') sources. **(B)** In the Bayesian causal inference model (adapted from Kording et al. (2007)), auditory (X_A) and visual (X_V) spatial signals are generated either by a common (C = 1, S_{AV}) or independent (C = 2) auditory (S_A) and visual (S_V) sources. **(C)** The 3 x 2 factorial model space manipulated (i) the implicit causal inference strategy involved in auditory spatial localization: model averaging (MA), model selection (MS) or probability matching (PM) and (ii) the explicit causal inference strategy involved in the common-source judgment: a fixed threshold of 0.5 ($CI_{Th-0.5}$) or a sampling strategy ($CI_{Sampling}$). The matrix shows the model evidences (i.e., Bayesian information criterions, BICs) of the 6 models relative to the worst model (larger = better). The bar plots show the family posterior probabilities of the three implicit causal inference model families (right) and the two explicit causal inference model families (top).

2.3 Materials and methods

Subjects

26 healthy subjects participated in the study after giving informed consent (16 female, mean age 25.8 years, range 23-37 years). All subjects had normal or corrected-to normal vision and reported normal hearing. The study was approved by the ethics committee of the University of Tübingen (protocol number 432 2007 BO1).

Stimuli

The visual stimulus was a cloud of 20 white dots (diameter: 0.43° visual angle; luminance 91 cd/m^2) sampled from a bivariate Gaussian presented on a dark grey background (luminance 62 cd/m^2, i.e., 47% contrast). The vertical standard deviation of the Gaussian was set to 5.4°. To manipulate the spatial reliability of the visual signal, the horizontal standard deviation was set to four levels: 0.1°, 5.4°, 10.8° or 16.2°. To manipulate the spatial location of the visual stimulus, the mean of the Gaussian was sampled from five possible locations along the azimuth (i.e., -10°, -3.3°, 0°, 3.3° or 10°). The auditory spatial signal was a burst of white noise. To create a virtual auditory spatial signal, the noise was convolved with spatially specific head-related transfer functions (HRTFs). The HRTFs were pseudo-individualized by matching subjects' head width, height and depth to the anthropometry of subjects in the CIPIC database (Algazi et al., 2001). HRTFs from the available locations in the database were interpolated to the desired locations of the auditory signal.

Experimental design and procedure

In a spatial ventriloquist paradigm (Fig. 2.1A), participants were presented with synchronous, yet spatially congruent or discrepant visual and auditory signals. On each trial, auditory and visual locations were independently sampled from five possible locations along the azimuth (i.e., -10°, -3.3°, 0°, 3.3° or 10°). In addition, we manipulated the reliability of the visual signal by setting the horizontal standard deviation of the Gaussian cloud to one of four possible levels (i.e., 0.1°, 5.4°, 10.8° or 16.2° STD). Hence, our experiment included 100 conditions arranged in a 5 (auditory location: S_A) x 5 (visual location: S_V) x 4 (visual reliability: $1/\sigma_V^2$) factorial design.

On each trial, synchronous auditory and visual spatial signals were presented for 32ms. Participants responded to two questions presented sequentially: First, participants localized the auditory spatial signal as accurately as possible by pushing one of five buttons that corresponded spatially to the stimulus locations (i.e., spatial localization). Second, participants decided whether the visual and auditory signals were generated by common or independent sources (i.e., common source judgment) and indicated their response via a two-choice key press. The time limit for both responses was 3s. The next trial started with a variable interval of 1-2.5s after participants had given their second button response. Throughout the experiment, participants fixated a cross (1.5° diameter) presented in the center of the screen.

The locations of the auditory and visual signals were randomized. The levels of visual reliability were presented either in blocks (55-85 trials per level of visual reliability) or varied according to a Markov chain (with a transition probability of 90% to stay on the same level of visual reliability and a 10% probability to change to an adjacent level). As the type of sequence did not influence the effects reported in this manuscript, we pooled over the two sequences and analyzed them together.

12 subjects participated in a longer version of the experiment including an additional level of visual reliability (21.6° STD). Data from this condition were excluded in the current analyses to have equivalent data sets from all 26 participants.

Overall, each participant completed 390-720 experimental trials. Prior to the main experiment, participants practiced the auditory localization task on 25 unisensory auditory trials, 25 audiovisual congruent trials with a single dot as the visual spatial signal and 15 trials with stimuli as in the main experiment.

Experimental setup

Audiovisual stimuli were presented using Psychtoolbox 3.09 (Brainard, 1997; Kleiner et al., 2007) (www.psychtoolbox.org) running under Matlab R2010b (MathWorks) on a Windows machine (Microsoft XP 2002 SP2). Auditory stimuli were presented at ~75 dB SPL using headphones (Sennheiser HD 555). Because visual stimuli required a large field of view, they were presented on a 30" LCD display (Dell UltraSharp 3007WFP). Participants were seated at a table in front of the screen in a darkened booth, resting their head on an adjustable chin rest. The viewing distance was 27cm resulting in a visual field of approx. 100°. Subjects indicated their responses using a standard keyboard. Subjects used the buttons {1,2,3,4,r} for spatial localization responses with their left hand and {9,0} for common-source judgments with their right hand.

Causal inference model

We employed a Bayesian causal inference (CI) model of audiovisual perception (Kording et al., 2007). On each trial, participants performed two tasks, an auditory localization and a common-source judgment. For each of the two tasks, we augmented the CI model with several decision strategies. For the implicit causal inference involved in the auditory localization task, we employed (i) model averaging, (ii) model selection and (iii) probability matching as previously described in Wozny et al. (2010). For the explicit causal inference involved in the common-source judgment, we used two decision functions that are described in detail below. By manipulating the decision functions for the spatial localization and the common-source judgment in a factorial fashion, we generated a 3 x 2 model space.

Details of the Bayesian generative model can be found in Kording et al. (2007). Briefly, we assume that a common (C = 1) or independent (C = 2) source is determined by sampling from a binomial distribution with the common-source prior $P(C = 1) = p_{\mathrm{common}}$ (Fig. 2.1B). For a common source, the 'true' location S_{AV} is drawn from the spatial prior distribution $N(\mu_P, \sigma_P)$. For two independent sources, the 'true' auditory (S_A) and visual (S_V) locations are drawn independently from this spatial prior distribution. For the spatial prior distribution, we assumed a central bias (i.e., $\mu_P = 0°$). We introduced sensory noise by independently drawing X_A and X_V from normal distributions centered on the true auditory (resp. visual) locations with parameters σ_A^2 (resp. σ_V^2). Thus, the generative model included the following free parameters: the common-source prior p_{common}, the spatial prior variance σ_P^2, the auditory variance σ_A^2 and the four visual variances σ_V^2 corresponding to the four visual reliability levels.

The probability of the underlying causal structure can be inferred by combining the common-source prior with the sensory evidence according to Bayes rule:

$$(1) \qquad p(C = 1|x_A, x_V) = \frac{p(x_A, x_V|C{=}1)p_{\mathrm{common}}}{p(x_A, x_V)}$$

In the case of a common source (C = 1; Fig. 2.1B left), the maximum a posteriori probability estimate of the auditory location is a reliability-weighted average of the auditory and visual estimates and the prior.

$$(2) \qquad \hat{S}_{A,C=1} = \frac{\frac{x_A}{\sigma_A{}^2} + \frac{x_V}{\sigma_V{}^2} + \frac{\mu_P}{\sigma_P{}^2}}{\frac{1}{\sigma_A{}^2} + \frac{1}{\sigma_V{}^2} + \frac{1}{\sigma_P{}^2}}$$

In the case of a separate-source inference (C = 2; Fig. 2.1B right), the estimate of the auditory signal location is independent from the visual spatial signal.

$$(3) \qquad \hat{S}_{A,C=2} = \frac{\frac{x_A}{\sigma_A{}^2} + \frac{\mu_P}{\sigma_P{}^2}}{\frac{1}{\sigma_A{}^2} + \frac{1}{\sigma_P{}^2}}$$

To provide a final estimate of the auditory location, the brain can combine the estimates under the two causal structures using various decision functions. In this study, we consider three decision functions for the implicit causal inference involved in the spatial localization task (for details see Wozny et al. (2010)): According to the 'model averaging' strategy, the brain combines the two auditory location estimates weighted in proportion to the posterior probability of their underlying causal structure.

$$\hat{S}_A = p(C=1|x_A, x_V)\ \hat{S}_{A,C=1} + (1 - p(C=1|x_A, x_V))\hat{S}_{A,C=2} \tag{4}$$

According to the 'model selection' strategy, the brain reports the spatial estimate selectively from the more likely causal structure.

$$\hat{S}_A = \begin{cases} \hat{S}_{A,C=1} & \text{if } p(C=1|x_A, x_V) > 0.5 \\ \hat{S}_{A,C=2} & \text{if } p(C=1|x_A, x_V) \leq 0.5 \end{cases} \tag{5}$$

According to 'probability matching', the brain reports the spatial estimate of one causal structure stochastically selected in proportion to its posterior probability.

$$\hat{S}_A = \begin{cases} \hat{S}_{A,C=1} & \text{if } p(C=1|x_A, x_V) > \alpha, & \alpha \sim U(0,1) \\ \hat{S}_{A,C=2} & \text{if } p(C=1|x_A, x_V) \leq \alpha, & \alpha \sim U(0,1) \end{cases} \tag{6}$$

Even though probability matching is sub-optimal, humans are known to use this strategy in a variety of cognitive tasks (e.g., Gaissmaier and Schooler (2008)). Further, a recent study suggested that human observers use probability matching in audiovisual spatial localization (Wozny et al., 2010).

We also considered two decision strategies for the explicit causal inference that is involved when generating a binary response (common source vs. independent sources) for the common-source judgment. First, we considered that subjects reported 'common source' when the posterior probability of a common source is greater than the threshold of 0.5 ('$CI_{Th\text{-}0.5}$').

$$\hat{C} = \begin{cases} 1 & \text{if } p(C=1|x_A, x_V) > 0.5 \\ 2 & \text{if } p(C=1|x_A, x_V) \leq 0.5 \end{cases} \tag{7}$$

Second, similar to the probability matching strategy described above for the spatial localization task, we considered that participants report 'common source' stochastically in proportion to the posterior probability of a common source ('$CI_{Sampling}$').

$$\hat{C} = \begin{cases} 1 & \text{if } p(C=1|x_A, x_V) > \alpha, & \alpha \sim U(0,1) \\ 2 & \text{if } p(C=1|x_A, x_V) \leq \alpha, & \alpha \sim U(0,1) \end{cases} \tag{8}$$

Factorially manipulating the decision functions for the spatial localization task and the common-source judgment, we generated a 3 x 2 space of six Bayesian CI models. We then fitted each of the six CI models jointly to the response data from the spatial localization and the common-source judgment tasks in a subject-specific fashion.

Fitting parameters of the six causal inference models

The predicted distributions of the auditory spatial estimates (i.e., $p(\hat{S}_A|S_A,S_{V,}1/\sigma_V^2)$) and the causal structure estimates (i.e., $p(\hat{C}|S_A,S_{V,}\ 1/\sigma_V^2)$) were obtained by marginalizing over the internal variables X_A and X_V. These distributions were generated by simulating X_A and X_V 1000 times for each of the 100 conditions and inferring $\hat{S}_A$ and $\hat{C}$ from equations (1)-(8). To link $p(\hat{S}_A|S_A,S_{V,}1/\sigma_V^2)$ to participants' auditory localization responses as discrete button responses, we assumed that participants selected the button that is closed to $\hat{S}_A$ and binned the data accordingly. Based on these predicted distributions, we computed the log likelihood of participants' auditory localization and causal judgment responses. Assuming independence of conditions and task responses, we summed the log likelihoods across conditions and across auditory localization and common-source judgment responses.

To obtain maximum likelihood estimates for the parameters of the models (p_{common}, σ_P, σ_A, σ_{V1} - σ_{V4} for each of the four levels of visual reliability), we used a non-linear simplex optimization algorithm as implemented in Matlab's fminsearch function (Matlab R2010b). This optimization algorithm was initialized with 200 different parameter settings that were defined based on a prior grid search. We report the results (across subjects' mean and standard error) from the parameter setting with the highest log likelihood across the 200 initializations (Tab. 2.1). This fitting procedure was applied individually to each participant's data set for each of the six CI models.

The model fit was assessed by the coefficient of determination (Nagelkerke, 1991). To identify the optimal model for explaining subjects' data, we compared the CI models using the Bayesian Information Criterion (BIC) as an approximation to the model evidence (Raftery, 1995). The BIC depends on both model complexity and model fit.

In addition, we investigated which decision strategy is most likely given the data separately for implicit causal inference during spatial localization and for explicit causal inference during common-source judgments. For this, we partitioned the model space into three (implicit causal inference) or two (explicit causal inference) model families according to the 3 x 2 factorial structure of our model space. Thus, we compared the three model families of model selection, model averaging and probability matching during the spatial localization task. Likewise, we compared the model families of fixed threshold at 0.5 and sampling procedure for the common-source judgment. The posterior probability of a model family is simply the sum of the posterior probabilities of each model within this family (Penny et al., 2010) (for implementational details see SPM8, www.fil.ion.ucl.ac.uk/spm, Friston et al. (1994)).

Comparing human responses to model predictions: Response indices

To inspect whether the most likely CI model can account qualitatively for participant's response profile, we show participants' responses and the predicted responses of the most likely CI model (Fig. 2.2-2.4). To enable a direct comparison, we processed and formed indices (e.g., the ventriloquist effect) of the model's predicted responses (1000 trials were simulated per condition) exactly as for the participants' responses (see below). For visualization and didactic purposes, we also present the predicted responses of a traditional forced-fusion model (Ernst and Banks, 2002; Alais and Burr, 2004) that is fitted selectively to the auditory localization data (Fig. 2.2B-D). Yet, the forced-fusion model cannot formally be compared to any of the CI models because it cannot be fitted to the common-source judgment data.

Specifically, we computed and presented the following response indices:
For the common-source judgment, we show the percentage of common-source judgments (Fig. 2.2A). For the spatial localization task, we present the absolute visual bias on the perceived auditory location, which is computed as the deviation of the responded auditory location from the true auditory location (i.e., $A_{Resp} - A_{Loc}$, Fig. 2.2B). Moreover, we show the ventriloquist effect (i.e., the relative visual bias on the perceived auditory location) computed as $VE = (A_{Resp} - A_{Loc}) / (V_{Loc} - A_{Loc})$ with A_{Resp} = mean auditory localization response for a given condition, A_{Loc} = auditory signal location and V_{Loc} = visual signal location (Fig. 2.2C, 3A-B). However, both the absolute and relative visual biases will be greater than zero, even when the visual signal has no influence on the auditory signal and vice versa. This is because participants predominantly make 'erroneous localization responses' towards more central positions in particular for extreme positions where they do not have the choice to respond to more eccentric positions. To account for these spatial response biases, we adjusted A_{Loc} and V_{Loc} with a linear regression approach across all congruent trials irrespective of the level of visual reliability in a subject-specific fashion. In other words, we replaced the true A_{Loc} and V_{Loc} in the crossmodal bias equations with the A_{Loc} and V_{Loc} predicted based on participants' responses during the congruent conditions. Based on simulation results, this adjustment procedure ensures that the crossmodal bias approximately measures the true underlying bias. Hence, the adjusted crossmodal bias reliably reflects the influence of a visual signal on auditory localization responses.

Finally, we evaluated the localization variability of the auditory localization responses as quantified by their variance (Fig. 2.2D, 2.3C-D):

$$(9) \qquad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(A_{Resp,\,i} - \bar{A}_{Resp})^2$$

Each of these response indices was computed for each of the 100 conditions in our 5 (auditory locations) x 5 (visual locations) x 4 (visual reliability levels) factorial design. We then reorganized these 100 conditions according to audiovisual spatial disparity and visual reliability (Fig. 2.2A-D). For this, we averaged the indices across all combinations of audiovisual locations at a particular level of spatial disparity and visual reliability (n.b. this averaging procedure is valid under the assumption that the visual bias is similar across different positions along the azimuth).

In addition, we analyzed the ventriloquist effect and localization variability as a function of common-source judgment by categorizing subjects' spatial localization responses according to whether participants responded common or separate source on those trials. If we treated the subjects' common-source judgment as an 'independent' factor, the factor induced an unbalanced distribution of trials across conditions, such that only few subjects had trials for all combinations of the factors spatial disparity, visual reliability and common-source judgment. Thus, for computing the ventriloquist effect, this analysis would have been limited to 13 subjects. Moreover, for computing the localization variability the analysis would have been limited even to only one single subject, as the computation of localization variability requires at least two trials per condition. When separating for common- vs. independent-source judgments, we therefore analyzed and presented the indices pooled either over the factor audiovisual disparity (Fig. 2.3A, C) or visual reliability (Fig. 2.3B, D). To ensure that the effects in the reliability x common-source design could be evaluated unconfounded by differences in disparity, we included only disparity levels that were present in all conditions for the remaining 4 (reliability) x 2 (common source) design in a particular subject. Likewise, when evaluating the effects in the disparity x common-source design, we included only those reliability levels that were present in all conditions for this design. This procedure enabled us to include full data sets from at least 25 subjects for the ventriloquist effect and the localization variability in both designs.

Model-free analysis of the causal judgments, ventriloquist effect and localization variability

The common-source judgments were characterized in terms of the percentage 'perceived common source' as a function of reliability and audiovisual spatial disparity. We then fitted Gaussian functions (i.e., a height, width and mean parameter) to the percentage 'perceived common source' as a function of the signed audiovisual disparity separately for each level of reliability (Fig. 2.2A). The effects of visual reliability on the height and width parameters were each assessed in a one-way repeated measures ANOVA.

The spatial localization responses were analyzed in terms of the relative audiovisual bias (i.e., ventriloquist effect) and the localization variability (i.e., variance). Both the ventriloquist effect and the localization variability were analyzed in separate visual reliability (4 levels) x spatial disparity repeated measures ANOVA. The factor spatial disparity had 5 levels for the localization variability, but only 4 levels for the ventriloquist effect as the computation of the ventriloquist effect requires a disparity greater zero.

We report Greenhouse-Geisser corrected p values and degrees of freedom. Effect sizes were reported as η^2.

2.4 Results

Comparison of the causal inference models

All six causal inference (CI) models were fitted jointly to participants' auditory localization and causal judgment responses and explained > 64% of the variance (R^2 > 64%; cf. Tab.

2.1). The smaller coefficient of determination results from the fact that the CI models in the current study only included 7 parameters to explain the variance across 100 conditions (compared to 4 parameters explaining 35 conditions in Kording et al. (2007) and Wozny et al. (2010)).

Table 2.1. Model parameters (mean ± SEM) and fit indices of the computational models.

Model	p_C	σ_P	σ_A	σ_{V1}	σ_{V2}	σ_{V3}	σ_{V4}	R^2	relBIC
MA & $CI_{Th\text{-}0.5}$	0.50±0.01	13.2±1.7	14.3±1.7	1.2±0.2	2.8±0.7	8.7±1.1	18.0±2.2	0.67±0.02	378.1
MS & $CI_{Th\text{-}0.5}$	0.51±0.01	10.2±1.4	12.2±1.1	3.2±0.2	7.2±1.3	9.0±1.0	14.3±1.9	0.64±0.03	0
PM & $CI_{Th\text{-}0.5}$	0.51±0.01	11.0±1.5	12.2±1.2	2.6±0.3	5.1±0.8	9.4±1.3	16.7±2.1	0.66±0.02	248.7
MA & $CI_{Sampling}$	0.62±0.02	13.6±2.1	12.7±1.8	1.6±0.3	2.7±0.3	7.8±1.0	16.9±2.1	0.67±0.02	316.4
MS & $CI_{Sampling}$	0.64±0.02	10.7±1.8	11.3±1.3	3.4±0.3	5.7±0.8	9.1±1.0	16.8±1.3	0.66±0.02	256.5
PM & $CI_{Sampling}$	0.63±0.02	11.1±1.9	10.0±0.9	2.7±0.2	4.0±0.3	6.9±0.5	17.1±2.2	0.66±0.02	243.5

Note: Models of the implicit causal inference strategy involved in auditory spatial localization: model averaging (MA), model selection (MS) or probability matching (PM). Models of the explicit causal inference strategy involved in the common-source judgment: a fixed threshold of 0.5 ($CI_{Th\text{-}0.5}$) or a sampling strategy ($CI_{Sampling}$). p_C = probability of the common-cause prior. σ_P = variance of the cue location prior (in °). σ_A = variance of the auditory percept (in °). σ_V = variance of the visual percept at different levels of visual reliability (1 = highest, 4 = lowest) (in °). R^2 = coefficient of determination (mean ± SEM). relBIC = Bayesian information criterion (BIC = LL - 0.5 M ln(N), LL = log likelihood, M = number of parameters, N = number of data points; BICs summed across sample) of a model relative to the worst model (larger = better).

Next, we identified the CI model that maximally accounts for participants' responses jointly during the auditory localization and the common-source judgment tasks by comparing the relative model evidence (BIC) of the CI models in our 3x2 model space (Fig. 2.1C). In the winning model, participants used the following decision strategies: For implicit causal inference in the auditory localization task, participants used model averaging as a decision strategy. Hence, they combined the spatial estimates under the two causal structures weighted by the posterior probabilities of each causal structure. For explicit causal inference during common-source judgments, participants reported 'common source' if the posterior probability was larger than an optimal threshold of 0.5 ('$CI_{Th\text{-}0.5}$'). The BIC difference between this model and the second best model was 61.7 which is generally considered as very strong evidence for the winning model (Raftery, 1995). Likewise, family inference (Penny et al., 2010) demonstrated the highest posterior probability for the model averaging strategy for the auditory localization task and the threshold ('$CI_{Th\text{-}0.5}$') decision strategy for the common-source judgment task (cf. Fig. 2.1C). In short, for both implicit and explicit causal inference, we did not observe evidence for a sampling strategy as was previously reported (Wozny et al., 2010).

Interestingly, the parameters for the visual variance of the winning CI model approximately matched the variance of the Gaussian cloud across the four visual reliability levels (Tab. 2.1). The auditory variance was comparable to the lowest visual variance. The common-source prior was approximately 0.5 indicating that participants a priori assumed that signals were equally likely to come from common or independent sources.

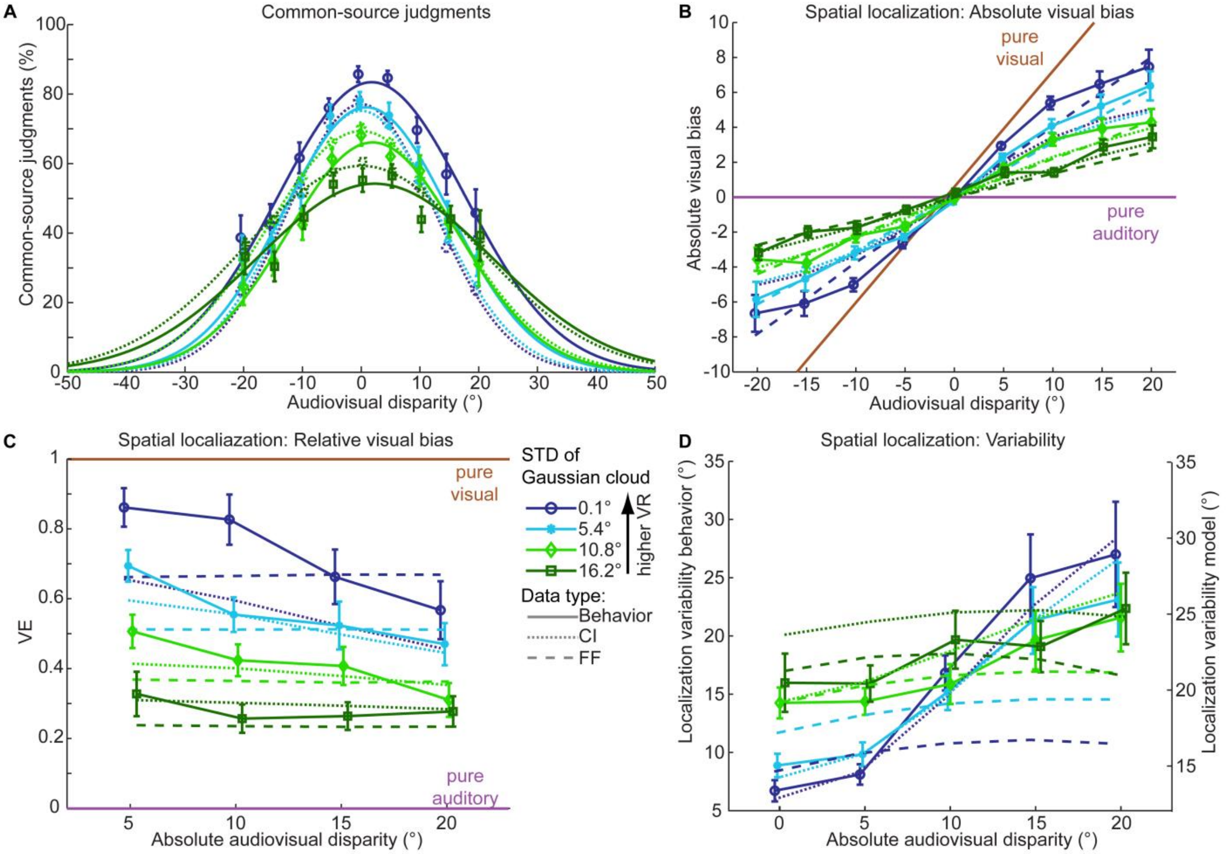


Figure 2.2. Behavioral responses and the models' predictions (pooled over common-source decisions). The figure panels show the behavioral data (mean ± SEM, solid lines) and the predictions of the winning causal inference model (CI, dotted lines) and the forced-fusion model (FF, dashed lines) as a function of visual reliability (color coded) and audiovisual disparity (shown along the x-axis). **(A)** Percentage of common-source judgments. **(B)** Absolute spatial visual bias, $A_{Resp} - A_{Loc}$. **(C)** Relative spatial visual bias (i.e., the ventriloquist effect VE = $(A_{Resp} - A_{Loc}) / (V_{Loc} - A_{Loc})$). In panels (B) and (C), the absolute and relative spatial visual bias are also shown for the case of pure visual or pure auditory influence for reference. **(D)** Localization variability of the behavioral and models' predicted responses (n.b. indicated on separate y-axes).

To further investigate whether the winning CI model qualitatively replicated participants' response profile, we compared participants' common-source judgments and auditory localization responses with the response predictions by the model. More specifically, we show the common-source judgments, the absolute and relative visual bias (i.e., ventriloquist effect) and the variability (i.e., variance) during the auditory localization task computed from participants' responses and the model's predicted responses.

Analysis of common-source judgments

The common-source judgments peaked at zero and decayed as a function of audiovisual disparity according to a Gaussian bell shaped function ($R^2 > 86\%$, explained variance averaged across the levels of visual reliability; Fig. 2.2A). A higher visual reliability significantly increased the height (effect of visual reliability on height parameter, $F_{2.5, 61.9} = 32.995$, $p < 0.001$, $\eta^2 = 0.569$) and marginally changed the width of the Gaussian (effect of visual reliability on width parameter, $F_{2.2, 54.2} = 2.606$, $p = 0.079$, $\eta^2 = 0.094$). The width of the Gaussian can be interpreted as an index for the width of the audiovisual integration window when it is judged explicitly in common-source judgments by participants. Our results demonstrate that participants were generally more likely to infer a common source at high relative to low visual reliability. Moreover, the slopes of the Gaussian functions were greater at high visual reliability indicating that high visual reliability rendered spatial disparity a more informative cue for discriminating between common source and independent sources.

Critically, the CI model qualitatively replicated, though slightly underestimated, the effect of visual reliability (cf. Fig. 2.2A). Thus, the model predicted fewer common-source judgments for high visual reliability and more frequent common-source judgments for low visual reliability.

Analysis of the visual bias and localization variability
- irrespective of common-source judgments

The visual influence on perceived auditory location was evaluated using the absolute visual bias (i.e., $A_{Resp} - A_{Loc,}$ Fig. 2.2B) and the relative visual bias also referred to as ventriloquist effect (i.e., $VE = (A_{Resp} - A_{Loc}) / (V_{Loc} - A_{Loc})$, Fig. 2.2C). Both indices are qualitatively in line with the predictions of Bayesian causal inference. Thus, the absolute visual bias increased non-linearly. This indicated that audiovisual integration breaks down when large spatial discrepancies render a common source unlikely. Likewise, for the relative visual bias (i.e., ventriloquist effect, Fig. 2.2C), we observed not only a main effect of visual reliability ($F_{1.6, 38.7} = 46.147$, $p < 0.001$, $\eta^2 = 0.649$) as predicted by forced-fusion models (Ernst and Banks, 2002; Alais and Burr, 2004), but also a main effect of absolute disparity ($F_{1.5, 37.8} = 21.339$, $p < 0.001$, $\eta^2 = 0.460$). Again as predicted by Bayesian causal inference, the ventriloquist effect is reduced for large spatial discrepancies when it is unlikely that the two signals come from a common source.

Critically, we also observed a significant interaction between visual reliability and spatial disparity (Fig. 2.2C; interaction effect of visual reliability and absolute disparity, $F_{5.5, 138.1} = 3.511$, $p = 0.004$, $\eta^2 = 0.123$). This interaction emerged because visual reliability changes the width and height of the audiovisual integration window. In other words, the shape of the Gaussian functions characterizing the common-source judgments (cf. Fig. 2.2A) indicates that less spatial disparity is needed for the brain to infer that audiovisual signals should be segregated at high visual reliabilities. These sharper audiovisual integration windows also make the ventriloquist effect decrease faster with spatial disparity when the visual signals are reliable.

The central benefit of multisensory integration is that it produces audiovisual estimates that are more reliable (i.e., less variable) (Ernst and Banks, 2002). Indeed, for small spatial disparities we observed that the auditory localization variability decreased with higher visual reliability (Fig. 2.2D). However, as predicted by Bayesian causal inference this reduction in localization variability was no longer observed for large spatial discrepancies indicating a breakdown of audiovisual integration (i.e., interaction effect of visual reliability and absolute disparity, $F_{4.6, 115.0} = 3.229$, $p = 0.011$, $\eta^2 = 0.114$; and a main effect of absolute disparity, $F_{1.8, 45.6} = 20.491$, $p < 0.001$, $\eta^2 = 0.450$). Note, however, that the CI model slightly overestimated the localization variability. For illustrational purposes, we therefore show the variance for participants' responses and for the model predictions using different axes to focus on their qualitative similarities.

Analysis of the visual bias, i.e. ventriloquist effect, and localization variability
- dependent on common-source judgments

Next, we investigated participants' auditory localization responses and the predictions of the CI model separately for trials on which participants inferred common or independent sources (Fig. 2.3). To illustrate how some of these effects on bias and localization variability emerge from splitting the localization response distributions according to the posterior common-source probability, we have also added figure 2.4 that shows the predicted distributions of the localization responses (along the y-axis) and posterior common-source probability (gray-tone coded) as a function of visual reliability (Fig. 2.4A) and spatial disparity (Fig. 2.4B).

As expected under Bayesian causal inference, we observed an overall larger ventriloquist effect that progressively increased with higher visual reliability when a common source was inferred (Fig. 2.3A). By contrast, when no common source was inferred, the ventriloquist effect was only negligibly influenced by visual reliability. Likewise, once the outcome of explicit causal inference was taken into account, the effect of spatial disparity (cf. Fig. 2.2C) was nearly abolished and the ventriloquist effect differed approximately by a constant when common and independent sources were inferred (Fig. 2.3B).

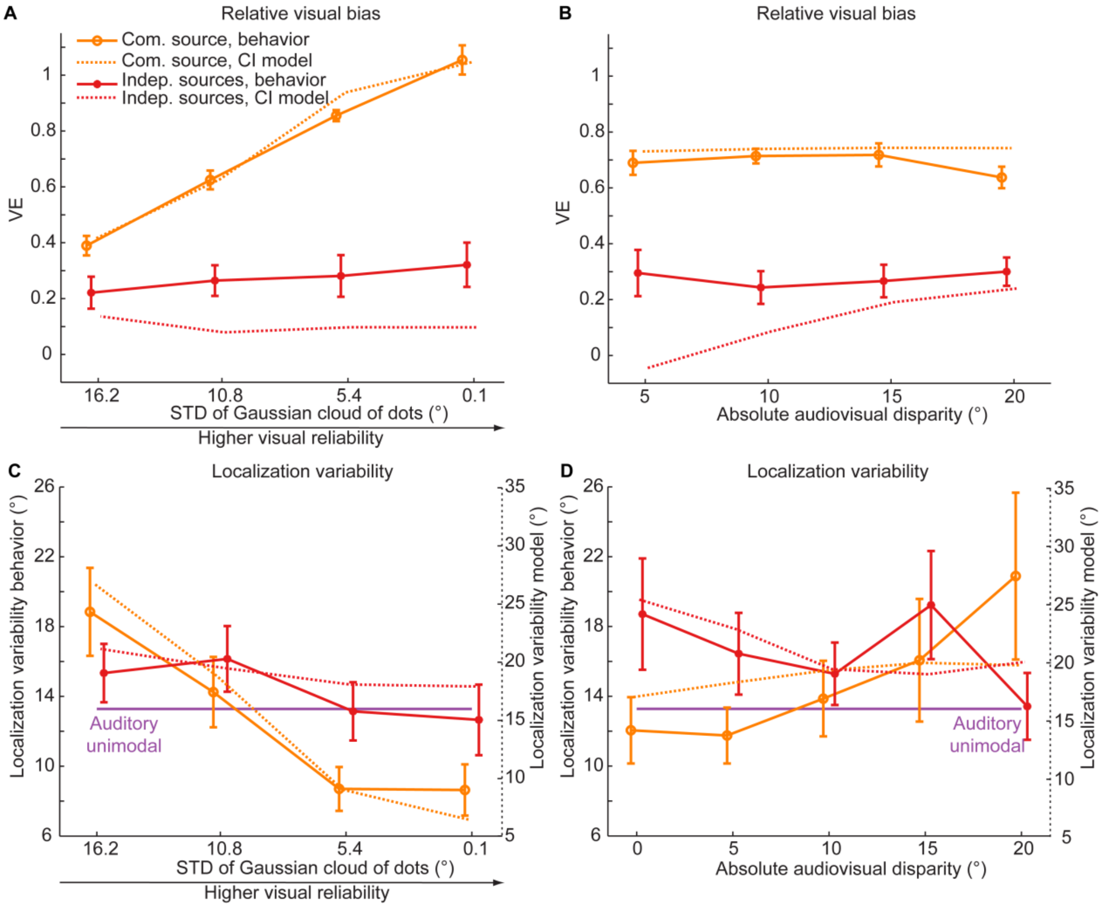


Figure 2.3. Behavioral responses and the model's predictions (separated according to common-source decisions). The figure panels show the behavioral data (mean ± SEM, solid lines) and the predictions of the winning causal inference model (CI, dotted lines) as a function of visual reliability (left, **A,C**) and audiovisual disparity (right, **B, D**). The ventriloquist effect (A, B) and localization variability (C, D, n.b. indicated in separate ordinate axes) are shown separately for trials where common or independent sources were inferred. Localization variability for unimodal auditory trials is shown as a solid line for reference.

While this is approximately in line with Bayesian causal inference, one would have predicted a repulsion effect for trials when separated sources were inferred. While a repulsion effect had indeed previously been shown for human localization responses, our study did not replicate this effect (Wallace et al., 2004; Kording et al., 2007). The reasons for these different behavioral response profiles are not clear. Potentially, a repulsion effect in our experiment was not observed because the visual stimulus was a cloud of dots rather than an LED flash or the sounds were delivered via headphones. Further, the previous study did not vary visual reliability across trials. Collectively, these experimental factors may have changed the consistency with which participants employ a decision threshold during the explicit common-source judgment resulting in a different separation of the spatial localization responses according to the common-source judgments.

Not surprisingly, the profile of auditory localization variability also depended on the outcome of participants' common-source judgment. As expected under the CI model, auditory localization variability was only negligibly influenced by visual reliability when participants inferred independent sources and segregated information. By contrast, the auditory localization variability was smaller than during unisensory conditions at least for high visual reliability when participants inferred a common source (Fig. 2.3C) and benefitted from audiovisual integration.

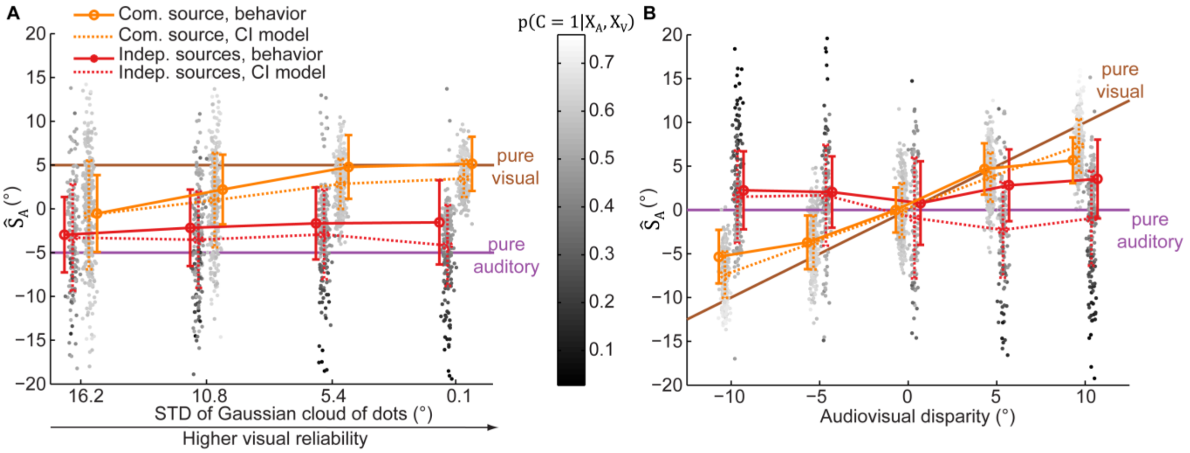


Figure 2.4. Distributions of auditory localization responses predicted by the Bayesian causal inference model. Distributions of auditory responses ($\hat{S}_{A,}$, along the y-axis) simulated by the winning Bayesian causal inference model (CI) as a function of visual reliability (left, A) and audiovisual disparity (right, B). The gray tone of each dot encodes the posterior common-source probability. For each level of reliability or disparity the dots are assigned to one of two clouds depending on whether the posterior probability of a common source ($p(C=1|x_A, x_V)$ is smaller than 0.5 (i.e., left cloud) or larger than 0.5 (i.e., right cloud). Further, the mean and the standard deviation of the predicted responses (dotted lines) and the observed behavioral responses (solid lines) are plotted. In (A), the visual and auditory signal locations are fixed at 5° and -5°, respectively (i.e., a constant spatial disparity of 10°). In (B), the visual reliability is fixed at 5.4° and the auditory signal location is fixed at 0°. For reference, the auditory responses in case of pure visual or pure auditory influence are shown as solid lines.

Likewise, the effect of spatial disparity on localization variability depended on the outcome of participants' common-source judgments (Fig. 2.3D). Interestingly, for both participants' and model's responses, the localization variability was decreased for small spatial disparities when a common source was inferred. Yet, it increased for spatial disparities when independent sources were inferred. This effect can be explained by the fact that participants infer independent sources predominantly when the observed visual signal is located far away from the auditory signal, either to its left or right. Thus, when no common source is inferred for small spatial discrepancies, the auditory localization responses come from a bimodal distribution leading to an increase in localization variability (see fig. 2.4B).

In conclusion, the behavioral response profile observed in the current study suggests that participants' explicit common-source judgment partially separates the spatial localization responses into two classes: When a common source is inferred, auditory localization responses conform approximately to predictions of the forced-fusion model. In other words, participants weight the sensory signals according to their reliability (Fig. 2.4A) in a linear fashion (Fig. 2.4B). By contrast, when no common source is inferred, participants responded predominantly based on the auditory signal approximately as predicted by a segregation model where signals are processed independently. Yet, while the explicit common-source judgment in our study provided only the binary response options 'common vs. separate' sources, the model averaging strategy weights the spatial estimates of the two causal structures by their continuous posterior probability. To relate explicit and implicit common-source judgments even more closely, a future study may therefore provide participants with a continuous response option (e.g., a rating scale) for the common-source judgment (Lewald and Guski, 2003).

2.5 Discussion

The current study investigated the decision strategies that observers use for inferring the causal structure of audiovisual spatial signals when probed implicitly in an auditory localization or explicitly in a causal judgment task. Given the critical role of sensory reliability in integration within (Jacobs, 1999; Knill and Saunders, 2003; Oruc et al., 2003) and across the senses (Yuille and Buelthoff, 1996; Ernst and Banks, 2002; Battaglia et al., 2003; Alais and Burr, 2004), we evaluated the causal inference model on psychophysics data that included multiple levels of visual reliability.

It is well established that sensory signals should only be integrated when they are close in time, space and structure (Welch and Warren, 1980; Slutsky and Recanzone, 2001; Lewald and Guski, 2003; Wallace et al., 2004; Roach et al., 2006). Recently, this problem has been framed within probabilistic Bayesian causal inference (Knill, 2007; Kording et al., 2007; Sato et al., 2007; Shams and Beierholm, 2010), where a response during implicit (i.e., in spatial localization) and explicit (i.e., common-source judgments) causal inference tasks can be formed based on several decision strategies (Wozny et al., 2010).

Our results show that human observers employ model averaging as a decision strategy for implicit causal inference in auditory localization. In other words, they obtain an auditory localization estimate by combining the spatial estimates under the two causal structures weighted by their posterior probabilities. The model averaging strategy minimizes the squared error of signal localizations and simultaneously accounts for the uncertainty of the underlying causal structure. By contrast, in a previous study the majority of participants used non-optimal probability matching for auditory localization (Wozny et al., 2010). These inconsistencies may arise from differences in the visual spatial signals (i.e., Gaussian cloud vs. LED) or the number of visual reliability levels (i.e., four vs. one) across the two experiments. Further, complex dual task effects (Stanovich and West, 2000; Stocker and Simoncelli, 2007) may explain the differences as the current design combined auditory localization and common-source judgment, while the previous study included auditory and visual localization tasks.

For explicit causal inference probed in the common-source judgment task, we observed that participants reported a common source if the common-source posterior probability was larger than 0.5. Thus, neither for implicit causal inference during spatial localization nor for explicit causal inference during common-source judgments did participants in our study employ sub-optimal sampling strategies where they selected each causal structure stochastically in proportion to its posterior probability.

Moving beyond previous modelling efforts (Kording et al., 2007; Sato et al., 2007; Beierholm et al., 2009; Wozny et al., 2010), we validated Bayesian causal inference models on a psychophysics data set that included several levels of visual reliabilities. This is critical, because according to the Bayesian causal inference model, sensory reliability influences multisensory integration via two distinct mechanisms: causal inference and reliability-weighted integration. Indeed, as expected under Bayesian causal inference, visual reliability sharpened the audiovisual integration window (Fig. 2.2A). Participants were better at discriminating whether sensory signals came from a common or two independent sources when the visual signals were highly reliable. As predicted by both forced fusion (Ernst and Banks, 2002; Alais and Burr, 2004) and causal inference models (Kording et al., 2007), high visual reliability also increased the influence of the visual signal on the perceived auditory location leading to a larger ventriloquist effect (Fig. 2.2C). Yet, in contradiction to the forced-fusion model, spatial ventriloquism broke down for greater spatial disparity when it is unlikely that audiovisual signals come from a common source. Moreover, we observed a significant interaction between reliability and spatial disparity. In other words, high visual reliability amplified the decay in ventriloquism with greater spatial disparity by sharpening the integration window. Likewise, the localization variability depended on both spatial disparity and visual reliability in an interactive fashion (Fig. 2.2D). In summary, visual reliability influenced the ventriloquist effect and localization variability via two interacting mechanisms: (i) sharpening of the integration window via causal inference and (ii) reliability-weighted integration in the case of a common source.

These two hierarchically organized mechanisms can be partially dissociated by separating localization responses depending on whether or not participants perceived a common source. Indeed, accounting for causal inference by separating trials according to participants' common-source judgments largely abolished the effect of spatial disparity on the ventriloquist effect and localization variability both for human responses and the predictions of the causal inference model (cf. Fig. 2.2C, D vs. Fig. 2.3B, D and 4B). Likewise, the effect of reliability on spatial ventriloquism and localization variability emerged predominantly when a common source was inferred (Fig. 2.3A, C and Fig. 2.4A). Collectively, these results suggest that reliability-weighted integration as a special case of multisensory integration is predicated on causal inferences. Yet, when separating localization responses according to the outcome of the common-source judgments, we still observed small effects of spatial discrepancy on spatial bias and localization variability. In particular, the localization variability increased for small spatial discrepancies when independent sources were inferred. As shown in figure 2.4, this surprising effect emerges because independent sources are inferred if the auditory percept is distant from the visual signal, either to the left or to the right, leading to a bimodal response distribution (Fig. 2.4B).

Research into the neural basis of multisensory integration has so far focused only on the special case of reliability-weighted integration under forced-fusion assumptions (Beauchamp et al., 2010; Helbig et al., 2012; Fetsch et al., 2013). For instance, very elegant neurophysiology work in macaque has demonstrated that single neurons integrate sensory inputs linearly weighted by their reliability (Morgan et al., 2008) in line with theories of probabilistic population coding (Ma et al., 2006). Furthermore, in a visuo-vestibular heading task decoding of neuronal activity in a dorsal medial superior temporal area (MSTd) mostly accounted for the sensory weights that the non-human primates employed at the behavioral level (Fetsch et al., 2012). It is currently unknown how the brain implements Bayesian causal inference during multisensory integration. Does it explicitly represent spatial estimates under forced-fusion and full-segregation assumptions as basic components of the causal inference model? Future fMRI in humans or neurophysiology studies in macaque are needed to address these questions.

In conclusion, the current study demonstrates that Bayesian causal inference is fundamental for multisensory integration in our natural uncertain environment. Sensory reliability critically shapes multisensory integration via two distinct mechanisms. First, it determines causal inference by sharpening the integration window. Second, it determines the relative weights of the sensory inputs in the integration process under the assumption of a common source.

2.6 Acknowledgments

We thank Ulrik Beierholm for helpful comments on a previous draft of the manuscript. The work was supported by the Max Planck Society and a grant of the European Research Council.

2.7 References

Alais D, Burr D (2004) The ventriloquist effect results from near-optimal bimodal integration. Curr Biol 14:257-262.

Algazi VR, Duda RO, Thompson DM, Avendano C (2001) The cipic hrtf database. In: Applications of Signal Processing to Audio and Acoustics, 2001 IEEE Workshop on the, pp 99-102: IEEE.

Battaglia PW, Jacobs RA, Aslin RN (2003) Bayesian integration of visual and auditory signals for spatial localization. J Opt Soc Am A Opt Image Sci Vis 20:1391-1397.

Beauchamp MS, Pasalar S, Ro T (2010) Neural substrates of reliability-weighted visual-tactile multisensory integration. Front Syst Neurosci 4:25.

Beierholm UR, Quartz SR, Shams L (2009) Bayesian priors are encoded independently from likelihoods in human multisensory perception. J Vis 9:23 21-29.

Brainard DH (1997) The psychophysics toolbox. Spatial vision 10:433-436.

Ernst MO, Banks MS (2002) Humans integrate visual and haptic information in a statistically optimal fashion. Nature 415:429-433.

Fetsch CR, Deangelis GC, Angelaki DE (2013) Bridging the gap between theories of sensory cue integration and the physiology of multisensory neurons. Nat Rev Neurosci 14:429-442.

Fetsch CR, Pouget A, DeAngelis GC, Angelaki DE (2012) Neural correlates of reliability-based cue weighting during multisensory integration. Nat Neurosci 15:146-154.

Friston KJ, Holmes AP, Worsley KJ, Poline JP, Frith CD, Frackowiak RSJ (1994) Statistical parametric maps in functional imaging: a general linear approach. Human brain mapping 2:189-210.

Gaissmaier W, Schooler LJ (2008) The smart potential behind probability matching. Cognition 109:416-422.

Helbig HB, Ernst MO, Ricciardi E, Pietrini P, Thielscher A, Mayer KM, Schultz J, Noppeney U (2012) The neural mechanisms of reliability weighted integration of shape information from vision and touch. Neuroimage 60:1063-1072.

Jacobs RA (1999) Optimal integration of texture and motion cues to depth. Vision Res 39:3621-3629.

Kleiner M, Brainard D, Pelli D, Ingling A, Murray R, Broussard C (2007) What's new in Psychtoolbox-3. Perception 36:1.1-16.

Knill DC (2007) Robust cue integration: a Bayesian model and evidence from cue-conflict studies with stereoscopic and figure cues to slant. J Vis 7:5 1-24.

Knill DC, Saunders JA (2003) Do humans optimally integrate stereo and texture information for judgments of surface slant? Vision Res 43:2539-2558.

Kording KP, Beierholm U, Ma WJ, Quartz S, Tenenbaum JB, Shams L (2007) Causal inference in multisensory perception. PLoS One 2:e943.

Lewald J, Guski R (2003) Cross-modal perceptual integration of spatially and temporally disparate auditory and visual stimuli. Brain Res Cogn Brain Res 16:468-478.

Ma WJ, Beck JM, Latham PE, Pouget A (2006) Bayesian inference with probabilistic population codes. Nat Neurosci 9:1432-1438.

Morgan ML, Deangelis GC, Angelaki DE (2008) Multisensory integration in macaque visual cortex depends on cue reliability. Neuron 59:662-673.

Nagelkerke NJ (1991) A note on a general definition of the coefficient of determination. Biometrika 78:691-692.

Oruc I, Maloney LT, Landy MS (2003) Weighted linear cue combination with possibly correlated error. Vision Res 43:2451-2468.

Penny WD, Stephan KE, Daunizeau J, Rosa MJ, Friston KJ, Schofield TM, Leff AP (2010) Comparing families of dynamic causal models. PLoS Comput Biol 6:e1000709.

Raftery AE (1995) Bayesian model selection in social research. Sociol Methodol 25:111-163.

Roach NW, Heron J, McGraw PV (2006) Resolving multisensory conflict: a strategy for balancing the costs and benefits of audio-visual integration. Proc Biol Sci 273:2159-2168.

Sato Y, Toyoizumi T, Aihara K (2007) Bayesian inference explains perception of unity and ventriloquism aftereffect: identification of common sources of audiovisual stimuli. Neural Comput 19:3335-3355.

Shams L, Beierholm UR (2010) Causal inference in perception. Trends Cogn Sci 14:425-432.

Slutsky DA, Recanzone GH (2001) Temporal and spatial dependency of the ventriloquism effect. Neuroreport 12:7-10.

Stanovich KE, West RF (2000) Individual differences in reasoning: implications for the rationality debate? Behav Brain Sci 23:645-665; discussion 665-726.

Stocker A, Simoncelli EP (2007) A Bayesian Model of Conditioned Perception. In: NIPS, pp 1409-1416.

Wallace MT, Roberson GE, Hairston WD, Stein BE, Vaughan JW, Schirillo JA (2004) Unifying multisensory signals across time and space. Exp Brain Res 158:252-258.

Welch RB, Warren DH (1980) Immediate perceptual response to intersensory discrepancy. Psychol Bull 88:638-667.

Wozny DR, Beierholm UR, Shams L (2010) Probability matching as a computational strategy used in perception. PLoS Comput Biol 6.

Yuille AL, Buelthoff HH (1996) Bayesian decision theory and psychophysics. New York: Cambridge University Press.

3 Cortical hierarchies jointly perform Bayesian causal inference for multisensory perception

3.1 Abstract

When faced with multisensory signals, the brain should only integrate the signals if they were caused by a common source to obtain a veridical percept of the environment. However, it is unknown whether and how cortical hierarchies take the signals' causal structure into account. Using Bayesian modelling and fMRI, we show that regions along auditory and visual spatial hierarchies jointly integrate audiovisual spatial signals according to their causal structure.

3.2 Introduction

To form a reliable percept of the multisensory environment, the brain integrates signals across the senses. However, it should integrate signals only when caused by a common source, but segregate those from different sources (Shams and Beierholm, 2010). Bayesian causal inference provides a rational strategy to arbitrate between information integration and segregation: In the case of a common source, signals should be integrated weighted by their sensory reliability (Ernst and Banks, 2002; Alais and Burr, 2004). In case of separate sources, they should be processed independently. Yet, in everyday life, the brain does not know the underlying causal structure, but needs to infer its probabilities from the sensory signals (Kording et al., 2007). A posterior signal estimate can then be obtained by averaging the estimates under the two causal structures weighted by the posterior probability of each causal structure (i.e., model averaging). Indeed, recent psychophysics research has demonstrated that human observers locate audiovisual signal sources according to Bayesian causal inference by combining the spatial estimates under the assumptions of common and separate sources weighted by their probabilities (Kording et al., 2007). Yet, despite recent evidence for a neural basis of reliability-weighted integration under a 'forced' assumption of a common source (Ma et al., 2006; Fetsch et al., 2012; Fetsch et al., 2013), the neural basis of Bayesian causal inference remains unknown. Thus, we combined Bayesian modeling and multivariate fMRI decoding to characterize how Bayesian causal inference is performed by the auditory (Tian et al., 2001) and visual (Mishkin et al., 1983) spatial cortical hierarchies.

3.3 Materials and methods

Participants

After giving written informed consent, six healthy volunteers without a history of neurological or psychiatric disorders (all university students or graduates; 2 female; mean age 28.8 years, range 22-36 years) participated in the fMRI study. All participants had normal or corrected-to normal vision and reported normal hearing. One participant was excluded because of excessive head motion (4.21 / 3.52 STD above the mean of the translational / rotational volume-wise head motion based on the included 5 participants). Note that the data from these 5 participants were also analyzed in chapter 4 and 5. The study was approved by the human research review committee of the University of Tuebingen.

Stimuli

The visual stimulus was a cloud of 20 white dots (diameter: 0.43° visual angle) sampled from a bivariate Gaussian with a vertical standard deviation of 2.5° and a horizontal standard deviation of 2° or 14° presented on a black background (i.e., 100% contrast). The auditory stimulus was a burst of white noise with a 5 ms on/off ramp. To create a virtual auditory spatial signal, the noise was convolved with spatially specific head-related transfer functions (HRTFs) thereby providing binaural (interaural time and amplitude differences) and monoaural spatial filtering signals. The HRTFs were pseudo-individualized by matching participants' head width, heights, depth and circumference to the anthropometry of participants in the CIPIC database (Algazi et al., 2001). HRTFs from the available locations in the database were interpolated to the desired location of the auditory signal.

Experimental design

In a spatial ventriloquist paradigm, participants were presented with synchronous, yet spatially congruent or discrepant visual and auditory signals (Fig. 3.1A). On each trial, visual and auditory locations were independently sampled from four possible locations along the azimuth (i.e., -10°, -3.3°, 3.3° or 10°) leading to four levels of spatial discrepancy (i.e., 0°, 6.6°, 13.3° or 20°). In addition, we manipulated the reliability of the visual signal by setting the horizontal standard deviation of the Gaussian cloud to 2° (high reliability) or 14° (low reliability) visual angle. In an inter-sensory selective-attention paradigm, participants either reported their auditory or visual perceived signal location and ignored signals of the other modality. Hence, the 4 x 4 x 2 x 2 factorial design manipulated (1) the location of the visual stimulus ({-10°, -3.3°, 3.3°, 10°}, i.e., the mean of the Gaussian) (2) the location of the auditory stimulus ({-10°, -3.3°, 3.3°, 10°}) (3) the reliability of the visual signal ({2°,14°}, STD of the Gaussian) and (4) task-relevance (auditory- / visual-selective report) (Fig. 3.1B). The design yielded 64 conditions.

On each trial, synchronous audiovisual spatial signals were presented for 50 ms followed by a variable inter-stimulus fixation interval from 1.75-2.75 s. Participants localized the signal in the task-relevant sensory modality as accurately as possible by pushing one of four spatially corresponding buttons. Throughout the experiment, they fixated a central cross (1.6° diameter).

To maximize design efficiency, stimuli and conditions were presented in a pseudorandomized fashion. Only the factor task-relevance was held constant within a session and counterbalanced across sessions. In each session, each of the 32 audiovisual spatial stimuli was presented exactly 11 times. 5.9% null-events were interspersed in the sequence of 352 stimuli per session. Each participant completed 20 sessions (10 auditory and 10 visual localization task; apart from one participant that performed 9 auditory and 11 visual localization sessions). Before the fMRI study, the participants completed one practice session outside the scanner.

Experimental setup

Audiovisual stimuli were presented using Psychtoolbox 3.09 (www.psychtoolbox.org) (Brainard, 1997) running under MATLAB R2010a (MathWorks). Auditory stimuli were presented at ~75 dB SPL using MR-compatible headphones (MR Confon). Visual stimuli were back-projected onto a Plexiglas screen using an LCoS projector (JVC DLA-SX21). Participants viewed the screen through an extra-wide mirror mounted on the MR head coil resulting in a horizontal visual field of approx. 76° at a viewing distance of 26 cm. Participants performed the localization task using an MR-compatible custom-built button device. Participants' eye movements and fixation were monitored by recording the participants' pupil location using an MR-compatible custom-build infrared camera (sampling rate 50 Hz) mounted in front of the participants' right eye and iView software 2.2.4 (SensoMotoric Instruments). Analyses of this data showed that participants did not commit to condition-related eye movements (cf. last paragraph of results in chapter 4.4 and supplemental tab. S4.1).

Bayesian causal inference model

To test the Bayesian causal inference model of audiovisual perception, we employed a generative model whose details can be found in Kording et al. (2007). The generative model (Fig. 3.1B) assumes that a common (C = 1) or independent (C=2) sources are determined by sampling from a binomial distribution with the common source prior P(C=1) = p_{common}. For a common source, the 'true' location S_{AV} is drawn from a spatial prior distribution N(μ_{AV}, σ_P). For two independent causes, the 'true' auditory (S_A) and visual (S_V) locations are drawn independently from this spatial prior distribution. For the spatial prior distribution, we assumed a central bias (i.e., $\mu = 0$). We introduced sensory noise by independently drawing X_A and X_V from normal distributions centered on the true auditory (resp. visual) locations with parameters σ_A^2 (resp. σ_V^2). Thus, the generative model included the following free parameters: the common source prior p_{common}, the spatial prior variance σ_P^2, the auditory variance σ_A^2 and the two visual variances σ_V^2 corresponding to the two visual reliability levels.

The probability of the underlying causal structure can be inferred by combining the common-source prior with the sensory evidence according to Bayes rule:

(1) $$p(C = 1|x_A, x_V) = \frac{p(x_A, x_V|C=1)p_{common}}{p(x_A, x_V)}$$

In the case of a common source (C = 1; Fig. 3.1B left), the optimal estimate of the audiovisual location is a reliability-weighted average of the auditory and visual percepts and the prior.

(2) $$\hat{S}_{AV,C=1} = \frac{\frac{x_A}{\sigma_A^2} + \frac{x_V}{\sigma_V^2} + \frac{\mu_P}{\sigma_P^2}}{\frac{1}{\sigma_A^2} + \frac{1}{\sigma_V^2} + \frac{1}{\sigma_P^2}}$$

In the case of independent sources (C = 2; Fig. 3.1B right), the optimal estimates of the auditory and visual signal locations (for the auditory and visual location report, respectively) are independent from signals of the ignored modality.

(3) $$\hat{S}_{A,C=2} = \frac{\frac{x_A}{\sigma_A^2} + \frac{\mu_P}{\sigma_P^2}}{\frac{1}{\sigma_A^2} + \frac{1}{\sigma_P^2}}, \quad \hat{S}_{V,C=2} = \frac{\frac{x_V}{\sigma_V^2} + \frac{\mu_P}{\sigma_P^2}}{\frac{1}{\sigma_V^2} + \frac{1}{\sigma_P^2}}$$

To provide a final estimate of the auditory and visual locations, the brain can combine the estimates under the two causal structures using various decision functions such as 'model averaging', 'model selection' and 'probability matching' (Wozny et al., 2010). Because the decision functions give highly correlated predictions and, therefore, are difficult to disentangle, we only present results using model averaging. According to the 'model averaging' strategy, the brain combines the integrated spatial estimate with the independent, task-relevant auditory or visual spatial estimates weighted in proportion to the posterior probability of their underlying causal structure.

(4) $$\hat{S}_A = p(C=1|x_A, x_V)\ \hat{S}_{AV,C=1} + (1 - p(C=1|x_A, x_V)\ )\hat{S}_{A,C=2}$$

(5) $$\hat{S}_V = p(C=1|x_A, x_V)\ \hat{S}_{AV,C=1} + (1 - p(C=1|x_A, x_V)\ )\hat{S}_{V,C=2}$$

Thus, full causal inference requires the brain to represent three spatial estimates ($\hat{S}_{AV,C=1}$, $\hat{S}_{A,C=2}$, $\hat{S}_{V,C=2}$) which are combined into a posterior estimate ($\hat{S}_A$ / $\hat{S}_V$, in dependence on which signal is task-relevant) by the posterior probability of the causal structure. However, a priori it is unknown whether an observer performs full causal inference (equation (4) and (5)) or only uses the component spatial estimates for signal localization: The observer could also fuse the audiovisual signals weighted by reliability in a forced fashion (equation (2)) or simply report the auditory (equation (3), left) or visual (equation (3), right) signals by segregating the signal which is irrelevant for the given localization task. Therefore, we compared the causal inference model against a forced fusion and a segregation model (Tab. 3.1).

To fit the causal inference, the forced fusion and the segregation models to participants' auditory and visual localization responses, we obtained the predicted distributions of the auditory spatial estimates (i.e., $p(\hat{S}_A|S_A,S_{V,}1/\sigma_V^2)$) and the visual spatial estimates (i.e., $p(\hat{S}_V|S_A,S_{V,}1/\sigma_V^2)$) by marginalizing over the internal variables X_A and X_V. These distributions were generated by simulating X_A and X_V 1000 times for each of the 64 conditions and inferring $\hat{S}_A$ and $\hat{S}_V$ from equations (1)-(5). To link $p(\hat{S}_A|S_A,S_{V,}1/\sigma_V^2)$ and $p(\hat{S}_V|S_A,S_{V,}1/\sigma_V^2)$ to participants' auditory or visual discrete localization responses, we assumed that participants selected the button that is closed to $\hat{S}_A$ or $\hat{S}_V$ and binned the data accordingly. Based on these predicted distributions, we computed the log likelihood of participants' auditory and visual localization responses. Assuming independence of conditions, we summed the log likelihoods across conditions.

To obtain maximum likelihood estimates for the parameters of the models (p_{common}, σ_P, σ_A, σ_{V1} - σ_{V2} for the two levels of visual reliability; the forced fusion and segregation models assumes p_{common} = 1 or = 0, respectively), we used a non-linear simplex optimization algorithm as implemented in Matlab's fminsearch function (Matlab R2010b). This optimization algorithm was initialized with 200 different parameter settings that were defined based on a prior grid search. We report the results (across participants' mean and standard error) from the parameter setting with the highest log likelihood across the 200 initializations (Tab. 3.1). This fitting procedure was applied individually to each participant's data set for each of the causal inference, the forced fusion und the unimodal segregation models.

The model fit was assessed by the coefficient of determination (Nagelkerke, 1991). To identify the optimal model for explaining participants' data, we compared the candidate models using the Bayesian Information Criterion (BIC) as an approximation for the model evidence (Raftery, 1995). The BIC depends on both model complexity and model fit. We performed Bayesian model selection (Stephan et al., 2009) as implemented in SPM8 (Friston et al., 1994) to obtain the exceedance probability for the candidate models (i.e., the probability that a given model is more likely than any other model given the data).

MRI data acquisition

A 3T Siemens Magnetom Trio MR scanner was used to acquire both T1-weighted anatomical images and T2*-weighted axial echoplanar images (EPI) with BOLD contrast (gradient echo, parallel imaging using GRAPPA with an acceleration factor of 2, TR = 2480 ms, TE = 40 ms, flip angle = 90°, FOV = 192×192 mm^2, image matrix 78×78, 42 transversal slices acquired interleaved in ascending direction, voxel size = 2.5×2.5×2.5 mm^3 + 0.25 mm interslice gap).

In total, 353 volumes times 20 sessions were acquired for the ventriloquist paradigm, 161 volumes times 2-4 sessions for the auditory localizer and 159 volumes times 10-16 sessions for the visual retinotopic localizer resulting in approximately 18 hours of scanning in total per participant assigned over 7-11 days. The first three volumes of each session were discarded to allow for T1 equilibration effects.

fMRI data analysis

Ventriloquist paradigm

The fMRI data were analyzed with SPM8 (www.fil.ion.ucl.ac.uk/spm) (Friston et al., 1994). Scans from each participant were corrected for slice timing, were realigned and unwarped to correct for head motion and spatially smoothed with a Gaussian kernel of 3 mm FWHM. The time series in each voxel was high-pass filtered to 1/128 Hz. All data were analyzed in native participant space. The fMRI experiment was modelled in an event-related fashion with regressors entering into the design matrix after convolving each event-related unit impulse with a canonical hemodynamic response function and its first temporal derivative. In addition to modelling the 32 conditions in our 4 (auditory locations) x 4 (visual locations) x 2 (visual reliability) factorial design, the general linear model included the realignment parameters as nuisance covariates to account for residual motion artefacts. The factor task-relevance (visual vs. auditory report) was modelled across sessions. The parameter estimates pertaining to the canonical hemodynamic response function defined the magnitude of the BOLD response to the audiovisual stimuli in each voxel. For the multivariate decoding analysis, we extracted the parameter estimates of the canonical hemodynamic response function for each condition and session from voxels of the regions of interest (= fMRI activation patterns) defined in separate auditory and retinotopic localizer experiments (see below). Each fMRI activation pattern for the 64 conditions in our 4x4x2x2 factorial design was based on 11 trials within a particular session. To avoid the effects of image-wide activity changes, each fMRI activation pattern was normalized to have mean zero and standard deviation one.

Decoding of spatial estimates

To investigate whether and how regions along the auditory and visual spatial hierarchy (defined below; cf. Fig. 3.1C) represent spatial estimates of the causal inference model, we used a multivariate approach to decode the estimates' values ($\hat{S}_{AV,C=1}$, $\hat{S}_{A,C=2}$, , $\hat{S}_{V,C=2}$, $\hat{S}_{A}$ / $\hat{S}_{V}$ combined across conditions of auditory and visual localization) from these regions. After fitting the causal inference model individually to behavioral localization responses (see above), the fitted model predicted the spatial estimates' values for each of the 64 conditions. For decoding, we used a linear support-vector regression model (SVR, as implemented in LIBSVM 3.14 (Chang and Lin, 2011). For each spatial estimate, we trained the SVR model to learn the mapping from the 64 fMRI activation patterns to the 64 spatial estimates' values from data of all but one session. The model then used this learnt mapping to decode the spatial estimates' values from the fMRI activation pattern from the remaining session. In a leave-one-out cross-validation scheme, the training-test procedure was repeated for all sessions.

To determine which spatial estimate of the causal inference model was most likely represented in a region, we compared the accuracies of decoding the four spatial estimates. For each estimate, the decoding accuracy was computed by predicting the decoded from the true spatial estimates using a linear regression. Decoded and true components were z standardized beforehand such that the parameter estimate represented the decoding accuracy as a correlation coefficient. Because spatial components are inherently highly correlated (up to r = 0.96 (mean across subjects) between $\hat{S}_{AV,C=1}$ and $\hat{S}_{V,C=2}$), it is difficult to select the component which a regions represents uniquely (i.e., a model selection problem). Therefore, we used a bootstrapping approach to evaluate whether a spatial estimate was more likely represented than any other estimate (i.e., the exceedance probability of a decoded spatial estimate) within a region (Burnham and Anderson, 2002). First, we bootstrapped decoding accuracies by resampling (N = 1000 times) with replacement the regression's residuals (Efron and Tibshirani, 1994) in each subject. We then determined for each bootstrap which spatial estimate had the highest average decoding accuracy (i.e., the mean of individual bootstrapped decoding accuracies after Fisher z transformation). The fraction of bootstraps in which a decoded spatial estimate had the highest decoding accuracy was the estimate's exceedance probability (Fig. 3.1D).

Auditory and visual retinotopic localizer

Auditory and visual retinotopic localizers were used to define regions of interest along the auditory and visual processing hierarchies in a participant-specific fashion. In the auditory localizer, participants were presented with brief bursts of white noise at -10° or 10° visual angle (duration 500 ms, stimulus onset asynchrony 2 s). In a one-back task, participants indicated via a key press when the spatial location of the current trial was different from the previous trial. 20 s blocks of auditory conditions (i.e., 20 trials) alternated with 13 s fixation periods. The auditory locations were presented in a pseudorandomized fashion to optimize design efficiency. Similar to the main experiment, the auditory localizer sessions were modelled in an event-related fashion with the onset vectors of left and right auditory stimuli being entered into the design matrix after convolution with the hemodynamic response function and its first temporal derivative. Auditory responsive regions were defined as voxels in superior temporal and Heschl's gyrus showing significant activations for auditory stimulation relative to fixation ($p < 0.05$, family-wise error corrected). Within these regions, we defined primary auditory cortex (A1) based on cytoarchitectonic probability maps (Eickhoff et al., 2005) and referred to the remainder (i.e., planum temporale and posterior superior temporal gyrus) as higher order auditory cortex (hA, see Fig. 3.1C).

Standard phase-encoded retinotopic mapping (Sereno et al., 1995) was used to define visual regions of interest. Participants viewed a checkerboard background flickering at 7.5 Hz through a rotating wedge aperture of 70° width (polar angle mapping) or an expanding/contracting ring (eccentricity mapping). The periodicity of the apertures was 42 s. Visual responses were modelled by entering a sine and cosine convolved with the hemodynamic response function as regressors in a general linear model. The preferred polar angle was determined as the phase lag for each voxel which is the angle between the parameter estimates for the sine and the cosine. The preferred phase lags for each voxel were projected on the reconstructed, inflated cortical surface using Freesurfer 5.1.0 (Dale et al., 1999). Visual regions V1-V3, V3AB and IPS0-IPS4 were defined as phase reversal in angular retinotopic maps. IPS0-4 were defined as contiguous, approximately rectangular regions based on phase reversals along the anatomical IPS (Swisher et al., 2007). For the decoding analyses, the auditory and visual regions were combined from the left and right hemispheres.

3.4 Results

To investigate how auditory and visual spatial cortical hierarchies perform causal inference, we presented 5 participants with synchronous auditory (white noise) and visual (cloud of dots) spatial signals independently sampled from 4 possible locations along the azimuth (i.e., -10°, -3.3°, 3.3° or 10°) whilst fMRI scanning (Fig. 3.1A). We manipulated the reliability of the cloud of dots (2° or 14° STD). Participants either selectively reported the visual or the auditory signal location. Thus, the four-factorial design (4 auditory locations x

4 visual locations x 2 levels of visual reliability x visual/auditory report) yielded 64 conditions.

When reporting the signals' location, it was a priori unclear whether participants performed causal inference by weighting the spatial estimates by the probability of the two potential causal structures (Fig. 3.1B, top). Alternatively, they integrated the audiovisual signal weighted by reliability under a forced assumption of a common source (Fig. 3.1B, C = 1, left), or they simply reported the auditory and visual signal locations by fully segregating the task-irrelevant signal under the assumption of separate sources (Fig. 3.1B, C = 2, right). After fitting the causal inference, forced fusion and segregation models to the participants' localization responses, we clearly found that the causal inference model provided a superior fit of the location reports (82.7% variance explained, exceedance probability of 0.953; Tab. 3.1).

Table 3.1. Model parameters (mean ± SEM) and fit indices of the three computational models.

Model	p_C	σ_P	σ_A	σ_{V1}	σ_{V2}	R^2	relBIC	EP
Causal inference	0.39+0.09	14.1+3.3	21.2+8.4	3.8+0.6	9.1+1.5	82.7+4.0	-7163.1+1044.0	0.9524
Forced fusion	-	14.4+1.6	14.3+2.0	6.5+0.5	10.7+0.7	60.7+3.4	-4293.4+388.0	0.0147
Segregation	-	13.1+2.7	24.1+9.9	4.1+0.7	7.5+0.9	79.1+4.3	-6707.4+1085.9	0.0329

Note: p_C = probability of the common-cause prior. σ_P = variance of the cue location prior (in °). σ_A = variance of the auditory percept (in °). σ_V = variance of the visual percept at different levels of visual reliability (1 = high, 2 = low) (in °). R^2 = coefficient of determination. relBIC = Bayesian information criterion (BIC = LL - 0.5 M ln(N), LL = log likelihood, M = number of parameters, N = number of data points; BICs summed across sample) of a model relative to the null model (larger = better). EP = exceedance probability, i.e. probability that a model is more likely than any other model.

Next, we investigated how regions along auditory and visual spatial cortical hierarchies (Fig. 3.1C) represent the spatial estimates of the causal inference model. First, we obtained four spatial estimates predicted by the individually fitted causal inference model for each of the 64 conditions (Fig. 3.1B, bottom): the reliability-weighted average under the assumption of a common source ($\hat{S}_{AV, C=1}$), the segregated unimodal estimates under the assumption of separate sources ($\hat{S}_{A, C=2}$, $\hat{S}_{V, C=2}$) and the final combined spatial estimate after averaging the reliability-weighted and the task-relevant unimodal estimate by the probability of common versus separate causes ($\hat{S}_A$/ $\hat{S}_V$, pooled over conditions of auditory and visual report). Using cross-validation, we trained a support vector regression model to decode these four spatial estimates from fMRI voxel response patterns in regions along the cortical hierarchies. We evaluated the decoding accuracy for each spatial estimate in terms of the correlation coefficient between the spatial estimate decoded from fMRI and predicted from the causal inference model. To determine the spatial estimate that is primarily encoded in a region, we next computed the exceedance probability that a correlation coefficient of one spatial estimate was greater than any of the other spatial estimates (Fig. 3.1D).

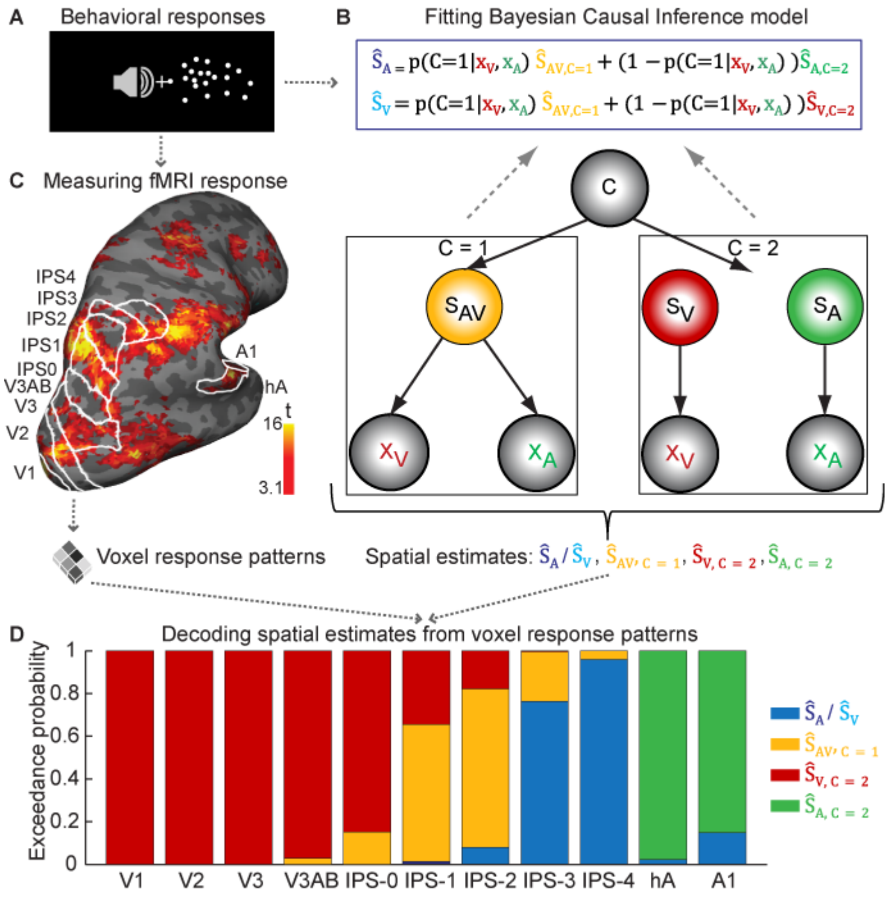


Figure 3.1. Stimuli, Bayesian causal inference model, cortical hierarchies and fMRI decoding results. (A) Participants were presented with audiovisual spatial signals from four possible locations at two levels of visual reliability. They either selectively localized the visual or the auditory signals (i.e., 64 experimental conditions in total) **(B)** The Bayesian causal inference model (Kording et al., 2007) was fit to participants' localization responses and then used to obtain spatial estimates: the unisensory auditory ($\hat{S}_{A,C=2}$) and visual ($\hat{S}_{V,C=2}$) estimates for separate signal sources (C = 2), the reliability-weighted average ($\hat{S}_{AV,C=1}$) for a common source (C = 1), and the final spatial estimate ($\hat{S}_A$, $\hat{S}_V$) that averages the task-relevant unisensory and the reliability-weighted common-source estimate weighted by the probability of their respective causal structures ($p(C = 1|x_A, x_V)$ or $(1 - p(C = 1|x_A, x_V))$), i.e. model averaging. **(C)** fMRI voxel response patterns were obtained from regions along the visual hierarchies, including lower visual regions (V1-3, V3AB) and intraparietal sulcus (IPS0-4), and auditory hierarchies, including primary (A1) and higher (hA) auditory cortex. **(D)** Exceedance probabilities index the belief that a given spatial estimate is more likely represented than any other spatial estimate within a region of interest. Exceedance probabilities were derived by comparing the decoding accuracies (i.e., quantified by correlation coefficient) of the spatial estimates from the voxel response patterns.

Thus, we found that Bayesian causal inference emerged along the auditory and visual hierarchies: Lower-level visual and auditory areas encoded auditory and visual estimates under the assumption of separate sources (i.e., information segregation). Posterior intraparietal sulcus (IPS1-2) represented the reliability-weighted average of the signals under a common-source assumption. Finally, anterior IPS (IPS3-4) represented the spatial estimates completing the causal inference: The region represented the average of the reliability-weighted and the task-relevant unimodal estimate weighted by the probability of common and separate causes, respectively.

3.5 Discussion

To our knowledge, this is the first demonstration that the computational operations underlying Bayesian causal inference are performed by the human brain in a hierarchical fashion. Critically, the brain explicitly encodes not only the spatial estimates under the assumption of full segregation (primary visual and auditory areas), but also under forced fusion (IPS1-2). These spatial estimates under causal structures of separate and common sources are then averaged into task-relevant auditory or visual estimate according to model averaging (IPS3-4).

Previous neurophysiological studies have shown that single neurons (Morgan et al., 2008) and population of neurons (Fetsch et al., 2012; Fetsch et al., 2013) implement forced-fusion reliability-weighted integration, presumably using a probabilistic population code (Ma et al., 2006). However, using the wide spatial coverage of the brain provided by fMRI, our results demonstrate that cortical hierarchies represent multiple multisensory spatial estimates which are jointly essential to integrate multisensory signals according to their causal structure. Future studies should investigate which brain regions explicitly compute the probability of common and separate sources which is the crucial quantity for balancing the signal estimates given their probabilistic causal structures. Further, it remains unknown which neural codes the brain uses to compute this causal-structure probability. Because this probability modulates reliability-weighted integration, our findings further call for an extension of the theory of probabilistic population codes (Ma et al., 2006) implementing reliability-weighted integration as well as signal segregation in dependence on the signals' causal structure.

Numerous studies demonstrated that large parts of neocortex have access to multisensory information (Ghazanfar and Schroeder, 2006; Driver and Noesselt, 2008). Even low-level sensory regions are influenced by crossmodal information from non-preferred modalities (Foxe et al., 2000; Molholm et al., 2002; Lewis and Noppeney, 2010). Such multisensory interactions arise via direct connections between unisensory regions (Falchier et al., 2002), top-down feedback from higher-order multisensory regions

(Macaluso and Driver, 2005) or via crossmodal thalamic input (Lakatos et al., 2007; Cappe et al., 2009). Importantly, we found that low-level visual and auditory regions represent their preferred spatial estimates, but this does not exclude the possibility of early crossmodal interactions (cf. results in chapter 4.4 and 5.4): Those interactions might rather adhere to the 'spatial principle' describing that multisensory interaction in neurons of superior colliculus are most pronounced if the multisensory signals jointly fall into the neuron's crossmodally registered receptive fields (Stein and Meredith, 1993). By contrast, our finding only suggests that the crossmodal interactions in low-level sensory regions are not in line with reliability-weighted averaging or full causal inference as found for IPS.

In conclusion, our study demonstrates that models of multisensory processes such as causal inference are essential to pinpoint specific multisensory processes along cortical hierarchies. Hence, our study provides a novel hierarchical perspective on multisensory integration in human neocortex.

3.6 Acknowledgments

This study was funded by the Max Planck Society and the European Research Council. We thank P. Ehses for advice on the fMRI sequence. We thank S. Schwarzkopf for providing stimulus scripts for retinotopic mapping (http://sampendu.wordpress.com/retinotopy-tutorial/).

3.7 References

Alais D, Burr D (2004) The ventriloquist effect results from near-optimal bimodal integration. Curr Biol 14:257-262.

Algazi VR, Duda RO, Thompson DM, Avendano C (2001) The cipic hrtf database. In: Applications of Signal Processing to Audio and Acoustics, 2001 IEEE Workshop on the, pp 99-102: IEEE.

Brainard DH (1997) The psychophysics toolbox. Spatial vision 10:433-436.

Burnham KP, Anderson DR (2002) Model selection and multi-model inference: a practical information-theoretic approach, 2nd Edition. New York: Springer.

Cappe C, Morel A, Barone P, Rouiller EM (2009) The thalamocortical projection systems in primate: an anatomical support for multisensory and sensorimotor interplay. Cereb Cortex 19:2025-2037.

Chang CC, Lin CJ (2011) LIBSVM: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology (TIST) 2:27.

Dale AM, Fischl B, Sereno MI (1999) Cortical surface-based analysis. I. Segmentation and surface reconstruction. Neuroimage 9:179-194.

Driver J, Noesselt T (2008) Multisensory interplay reveals crossmodal influences on 'sensory-specific' brain regions, neural responses, and judgments. Neuron 57:11-23.

Efron B, Tibshirani RJ (1994) An introduction to the bootstrap. London: Chapmann and Hall.

Eickhoff SB, Stephan KE, Mohlberg H, Grefkes C, Fink GR, Amunts K, Zilles K (2005) A new SPM toolbox for combining probabilistic cytoarchitectonic maps and functional imaging data. Neuroimage 25:1325-1335.

Ernst MO, Banks MS (2002) Humans integrate visual and haptic information in a statistically optimal fashion. Nature 415:429-433.

Falchier A, Clavagnier S, Barone P, Kennedy H (2002) Anatomical evidence of multimodal integration in primate striate cortex. J Neurosci 22:5749-5759.

Fetsch CR, Deangelis GC, Angelaki DE (2013) Bridging the gap between theories of sensory cue integration and the physiology of multisensory neurons. Nat Rev Neurosci 14:429-442.

Fetsch CR, Pouget A, DeAngelis GC, Angelaki DE (2012) Neural correlates of reliability-based cue weighting during multisensory integration. Nat Neurosci 15:146-154.

Foxe JJ, Morocz IA, Murray MM, Higgins BA, Javitt DC, Schroeder CE (2000) Multisensory auditory-somatosensory interactions in early cortical processing revealed by high-density electrical mapping. Brain Res Cogn Brain Res 10:77-83.

Friston KJ, Holmes AP, Worsley KJ, Poline JP, Frith CD, Frackowiak RSJ (1994) Statistical parametric maps in functional imaging: a general linear approach. Human brain mapping 2:189-210.

Ghazanfar AA, Schroeder CE (2006) Is neocortex essentially multisensory? Trends Cogn Sci 10:278-285.

Kording KP, Beierholm U, Ma WJ, Quartz S, Tenenbaum JB, Shams L (2007) Causal inference in multisensory perception. PLoS One 2:e943.

Lakatos P, Chen CM, O'Connell MN, Mills A, Schroeder CE (2007) Neuronal oscillations and multisensory interaction in primary auditory cortex. Neuron 53:279-292.

Lewis R, Noppeney U (2010) Audiovisual synchrony improves motion discrimination via enhanced connectivity between early visual and auditory areas. J Neurosci 30:12329-12339.

Ma WJ, Beck JM, Latham PE, Pouget A (2006) Bayesian inference with probabilistic population codes. Nat Neurosci 9:1432-1438.

Macaluso E, Driver J (2005) Multisensory spatial interactions: a window onto functional integration in the human brain. Trends Neurosci 28:264-271.

Mishkin M, Ungerleider LG, Macko KA (1983) Object vision and spatial vision: two cortical pathways. Trends in neurosciences 6:414-417.

Molholm S, Ritter W, Murray MM, Javitt DC, Schroeder CE, Foxe JJ (2002) Multisensory auditory-visual interactions during early sensory processing in humans: a high-density electrical mapping study. Brain Res Cogn Brain Res 14:115-128.

Morgan ML, Deangelis GC, Angelaki DE (2008) Multisensory integration in macaque visual cortex depends on cue reliability. Neuron 59:662-673.

Nagelkerke NJ (1991) A note on a general definition of the coefficient of determination. Biometrika 78:691-692.

Raftery AE (1995) Bayesian model selection in social research. Sociol Methodol 25:111-163.

Sereno MI, Dale AM, Reppas JB, Kwong KK, Belliveau JW, Brady TJ, Rosen BR, Tootell RB (1995) Borders of multiple visual areas in humans revealed by functional magnetic resonance imaging. Science 268:889-893.

Shams L, Beierholm UR (2010) Causal inference in perception. Trends Cogn Sci 14:425-432.

Stein BE, Meredith MA (1993) The merging of the senses. Cambridge, MA: The MIT Press.

Stephan KE, Penny WD, Daunizeau J, Moran RJ, Friston KJ (2009) Bayesian model selection for group studies. Neuroimage 46:1004-1017.

Swisher JD, Halko MA, Merabet LB, McMains SA, Somers DC (2007) Visual topography of human intraparietal sulcus. J Neurosci 27:5326-5337.

Tian B, Reser D, Durham A, Kustov A, Rauschecker JP (2001) Functional specialization in rhesus monkey auditory cortex. Science 292:290-293.

Wozny DR, Beierholm UR, Shams L (2010) Probability matching as a computational strategy used in perception. PLoS Comput Biol 6.

4 To integrate, or not to integrate: Causal inference in primary sensory and association cortices during multisensory perception

4.1 Abstract

To form a reliable percept of the multisensory environment, the brain needs to integrate signals caused by a common source, but segregate those from different sources. Bayesian causal inference provides a rational strategy to arbitrate between information integration and segregation. Yet, its neural basis is unknown. In this functional magnetic resonance imaging (fMRI) study, participants localized audiovisual signals that varied in spatial discrepancy and visual reliability. While multivariate fMRI decoding revealed crossmodal influences already in primary sensory areas, only higher-order intraparietal sulci (IPS) integrated audiovisual signals weighted by their bottom-up sensory reliability and top-down task-relevance. Critically, audiovisual integration was attenuated for large spatial discrepancies when it is unlikely that audiovisual signals originate from a common source. In line with the principles of Bayesian causal inference our results demonstrate that IPS integrates audiovisual signals into spatial priority maps by taking into account the probabilities of the environmental causal structures.

4.2 Introduction

Information integration and segregation is a fundamental task facing the brain in numerous perceptual and cognitive contexts (Shams and Beierholm, 2010). Most prominently, in our natural environment our senses are constantly bombarded with many different signals that provide uncertain information about the world. To form a reliable representation of the environment, the brain is challenged to integrate noisy signals originating from a common source and segregate those from different sources (Kording et al., 2007; Shams and Beierholm, 2010).

Behaviorally, humans typically integrate sensory signals weighted by their relative reliability when they are close in time (Parise et al., 2012), space (Alais and Burr, 2004) and structure (Ernst and Banks, 2002). Recent elegant neurophysiological studies in macaque (Fetsch et al., 2013) have characterized how this reliability-weighted integration is implemented by single neurons and neuronal populations during a visual-vestibular heading task (Morgan et al., 2008; Fetsch et al., 2012). Consistent with theories of probabilistic population codes (Ma et al., 2006), they demonstrated that the dorsal medial superior temporal area combines visual and vestibular inputs linearly in proportion to their reliability.

Reliability-weighted integration is statistically optimal for signals coming from a common source. Yet, it fails to capture more natural situations where the brain is confronted with many signals that may or may not arise from a common source (Shams and Beierholm, 2010). Indeed, it is well-established that information integration breaks down for conflicting sensory signals (Welch and Warren, 1980; Stein and Meredith, 1993) that are unlikely to emanate from a common source. Most prominently, the ventriloquist illusion which illustrates how sensory signals are integrated into spatial representations is strongly modulated by the spatial discrepancy between the sensory signals (Wallace et al., 2004). For small spatial discrepancies, the perceived auditory location shifts towards the visual location and vice versa depending on the relative sensory reliabilities (Alais and Burr, 2004); for large spatial discrepancies, these audiovisual biases and integration processes are greatly attenuated (Bertelson and Radeau, 1981).

Recent behavioral studies have demonstrated that Bayesian causal inference can well account for this behavioral profile by explicitly modelling the potential causal structures of the sensory inputs (Kording et al., 2007). Under the assumption of a common source auditory and visual spatial estimates are combined weighted according to their reliabilities; under the hypothesis of different sources, they are treated independently. Critically, on a particular instance the brain does not know the underlying causal structures. It needs to infer their probabilities from the sensory inputs based on audiovisual correspondences such as spatial discrepancy. The brain is then thought to compute a final estimate of the auditory (or visual) signal location by combining the spatial estimates under the two causal assumptions (i.e., common vs. independent sources) weighted by their probability (Kording et al., 2007; Wozny et al., 2010). Causal inference thus provides the brain with a rational strategy to arbitrate between information integration and segregation by taking into account the probabilities of the underlying causal structures.

Using multivariate pattern decoding this fMRI study investigated how the human brain integrates audiovisual signals into spatial representations depending on the causal structure of the world. In a spatial ventriloquist paradigm, we presented participants with audiovisual signals while manipulating bottom-up sensory reliability, top-down task-relevance (i.e., visual vs. auditory report) and spatial discrepancy of the audiovisual signals. Our results demonstrate that the intraparietal sulcus (IPS) integrates audiovisual signals into spatial priority maps in line with the principles of Bayesian causal inference.

4.3 Materials and methods

Participants

After giving written informed consent, six healthy volunteers without a history of neurological or psychiatric disorders (all university students or graduates; 2 female; mean age 28.8 years, range 22-36 years) participated in the fMRI study. All participants had normal or corrected-to normal vision and reported normal hearing. One participant was excluded because of excessive head motion (4.206 / 3.518 STD above the mean of the translational / rotational volume-wise head motion based on the included 5 participants). Note that the data from these 5 participants were also analyzed in chapter 3 and 5. The study was approved by the human research review committee of the University of Tuebingen.

Stimuli

The visual stimulus was a cloud of 20 white dots (diameter: 0.43° visual angle) sampled from a bivariate Gaussian with a vertical standard deviation of 2.5° and a horizontal standard deviation of 2° or 14° presented on a black background (i.e., 100% contrast). The auditory stimulus was a burst of white noise with a 5 ms on/off ramp. To create a virtual auditory spatial signal, the noise was convolved with spatially specific head-related transfer functions (HRTFs) thereby providing binaural (interaural time and amplitude differences) and monoaural spatial filtering signals. The HRTFs were pseudo-individualized by matching participants' head width, heights, depth and circumference to the anthropometry of participants in the CIPIC database (Algazi et al., 2001). HRTFs from the available locations in the database were interpolated to the desired location of the auditory signal.

Experimental design

In a spatial ventriloquist paradigm, participants were presented with synchronous, yet spatially congruent or discrepant visual and auditory signals (Fig. 4.1A). On each trial, visual and auditory locations were independently sampled from four possible locations along the azimuth (i.e., -10°, -3.3°, 3.3° or 10°) leading to four levels of spatial discrepancy (i.e., 0°, 6.6°, 13.3° or 20°). In addition, we manipulated the reliability of the visual signal by setting the horizontal standard deviation of the Gaussian cloud to 2° (high reliability) or 14° (low reliability) visual angle. In an inter-sensory selective-attention paradigm, participants either reported their auditory or visual perceived stimulus location. Hence, the 4 x 4 x 2 x 2 factorial design manipulated (1) the location of the visual stimulus ({-10°, -3.3°, 3.3°, 10°}, i.e., the mean of the Gaussian) (2) the location of the auditory stimulus ({-

10°, -3.3°, 3.3°, 10°}) (3) the reliability of the visual signal ({2°,14°}, STD of the Gaussian) and (4) task-relevance (auditory- / visual-selective report) (Fig. 4.1B).

On each trial, synchronous audiovisual spatial signals were presented for 50 ms followed by a variable inter-stimulus fixation interval from 1.75-2.75 s. Participants localized the signal in the task-relevant sensory modality as accurately as possible by pushing one of four spatially corresponding buttons. Throughout the experiment, they fixated a central cross (1.6° diameter).

To maximize design efficiency, stimuli and conditions were presented in a pseudorandomized fashion. Only the factor task-relevance was held constant within a session and counterbalanced across sessions. In each session, each of the 32 audiovisual spatial stimuli was presented exactly 11 times. 5.9% null-events were interspersed in the sequence of 352 stimuli per session. Each participant completed 20 sessions (10 auditory and 10 visual localization task; apart from one participant that performed 9 auditory and 11 visual localization sessions). Before the fMRI study, the participants completed one practice session outside the scanner.

The number of sessions of the main experiment (i.e., 10 sessions x 2 task-contexts = 20 sessions) was determined based on a prior independent pilot study with one single subject that participated in 33 sessions in total including 17 scanning sessions for the auditory localization task. Computing the decoding performance for an increasing number of sessions demonstrated that reliable decoding performance was obtained approximately for ≥ 10 sessions (Supplemental fig. S4.2 and figure legend). Moreover, as the effect of visual reliability on audiovisual reweighting (cf. Fig. 4.3) was large (Cohen's d = 0.829 in IPS0 in the pilot participant), we decided to scan a small sample of six participants extensively in 20 sessions.

Experimental setup

Audiovisual stimuli were presented using Psychtoolbox 3.09 (www.psychtoolbox.org) (Brainard, 1997) running under MATLAB R2010a (MathWorks). Auditory stimuli were presented at ~75 dB SPL using MR-compatible headphones (MR Confon). Visual stimuli were back-projected onto a Plexiglas screen using an LCoS projector (JVC DLA-SX21). Participants viewed the screen through an extra-wide mirror mounted on the MR head coil resulting in a horizontal visual field of approx. 76° at a viewing distance of 26 cm. Participants performed the localization task using an MR-compatible custom-built button device. Participants' eye movements and fixation were monitored by recording the participants' pupil location using an MR-compatible custom-build infrared camera (sampling rate 50 Hz) mounted in front of the participants' right eye and iView software 2.2.4 (SensoMotoric Instruments).

Behavioral data

We quantified the relative influence of the visual and the auditory signal on the reported location as crossmodal bias CMB = ($L_{response}$ − $L_{Auditory}$) / (L_{Visual} − $L_{Auditory}$) for the incongruent trials. $L_{response}$ denotes the reported location and L_{Visual} (or $L_{Auditory}$) the location of the visual (or auditory) signal. To accommodate the participant-specific central response bias, we adjusted L_{Visual} and $L_{Auditory}$ using a linear regression based on the congruent trials alone. In these regression models, the reported visual (or auditory) locations were predicted by the true position of the visual (or auditory) locations including all congruent visual (or auditory) report trials irrespective of sensory reliability. The predicted visual (or auditory) locations were entered as the L_{Visual} (or $L_{Auditory}$) in the formula to compute the crossmodal bias.

The CMB was analyzed using a two (task-relevance: auditory vs. visual report) x two (visual reliability: high vs. low) x two (spatial discrepancy: small (≤ 6.6°) vs. large (> 6.6°)) repeated measure ANOVA. To obtain more efficient and balanced estimates, we pooled over two levels of discrepancy. CMBs were normally distributed across participants ($p \geq 0.797$ in Kolmogorov-Smirnov tests in each of the 2 x 2 x 2 conditions).

MRI data acquisition

A 3T Siemens Magnetom Trio MR scanner was used to acquire both T1-weighted anatomical images and T2*-weighted axial echoplanar images (EPI) with BOLD contrast (gradient echo, parallel imaging using GRAPPA with an acceleration factor of 2, TR = 2480 ms, TE = 40 ms, flip angle = 90°, FOV = 192×192 mm^2, image matrix 78×78, 42 transversal slices acquired interleaved in ascending direction, voxel size = 2.5×2.5×2.5 mm^3 + 0.25 mm interslice gap).

In total, 353 volumes times 20 sessions were acquired for the ventriloquist paradigm, 161 volumes times 2-4 sessions for the auditory localizer and 159 volumes times 10-16 sessions for the visual retinotopic localizer resulting in approximately 18 hours of scanning in total per participant assigned over 7-11 days. The first three volumes of each session were discarded to allow for T1 equilibration effects.

fMRI data analysis

Ventriloquist paradigm

The fMRI data were analyzed with SPM8 (www.fil.ion.ucl.ac.uk/spm) (Friston et al., 1994). Scans from each participant were corrected for slice timing, were realigned and unwarped to correct for head motion and spatially smoothed with a Gaussian kernel of 3 mm FWHM. The time series in each voxel was high-pass filtered to 1/128 Hz. All data were analyzed in native participant space. The fMRI experiment was modelled in an event-related fashion with regressors entering into the design matrix after convolving each event-related unit impulse with a canonical hemodynamic response function and its first temporal derivative. In addition to modelling the 32 conditions in our 4 (auditory locations) x 4 (visual locations) x 2 (visual reliability) factorial design, the general linear model included the realignment parameters as nuisance covariates to account for residual motion artefacts. The factor task-relevance (visual vs. auditory report) was modelled across sessions. The parameter estimates pertaining to the canonical hemodynamic response function defined the magnitude of the BOLD response to the audiovisual stimuli in each voxel. For the multivariate decoding analysis, we extracted the parameter estimates of the canonical hemodynamic response function for each condition and session from voxels of the regions of interest (= fMRI activation vectors) defined in separate auditory and retinotopic localizer experiments (see below). Each fMRI activation vector for the 64 conditions in our 4x4x2x2 factorial design was based on 11 trials within a particular session. To avoid the effects of image-wide activity changes, each fMRI activation vector was normalized to have mean zero and standard deviation one.

For the multivariate decoding analysis, we used a linear support-vector regression model (SVR, as implemented in LIBSVM 3.14 (Chang and Lin, 2011) to determine how regions of interest in the visual and auditory processing hierarchies integrate auditory and visual signals into spatial representations. We trained the SVR model to learn the mapping from the fMRI activation vectors to the external spatial locations based on the audiovisually *congruent* conditions (including conditions of auditory and visual report) from all but one session. This learnt mapping from activation pattern to external spatial location was then used to decode the spatial location from the fMRI activation vectors of the spatially congruent and incongruent audiovisual conditions of the remaining session (see fig. 4.2A). In a leave-one-out cross-validation scheme, the training-test procedure was repeated for all sessions.

By this procedure, the decoded spatial locations of the spatially incongruent conditions provide information about how a brain region combines visual and auditory spatial signals into spatial representations. To quantify the influence of the auditory and visual signals on the decoded spatial location, we used a linear regression approach (Fig. 4.2B) where we predicted the decoded spatial location by the true auditory and true visual signal locations. Based on the principles of Bayesian causal inference, we expected that the influence of the true auditory and true visual location on the decoded spatial position would depend on three factors: (i) signal reliability, (ii) task relevance and (iii) audiovisual spatial discrepancy. Hence, we entered the true auditory and visual locations separately as explanatory variables into this regression model for each condition in a two (visual reliability: high vs. low) x two (task-relevance: auditory vs. visual report) x two (spatial discrepancy: ≤ 6.6° vs. > 6.6°) factorial design (i.e., 8 (conditions) x 2 (true auditory or visual spatial locations) = 16 regressors in total). The auditory ($ß_A$) and visual ($ß_V$) parameter estimates quantified the influence of auditory and visual signals on the decoded spatial location for a particular condition. To increase the efficiency of the parameter estimates in higher order association areas, we pooled the decoded spatial locations in IPS0-2 and IPS3-4 in the regression model (see (Liu et al., 2011) for a similar approach).

For each condition in the two (visual reliability: high vs. low) x two (task-relevance: auditory vs. visual report) x two (spatial discrepancy: ≤ 6.6° vs. > 6.6°) factorial design, we computed the relative audiovisual weight as the angle w_{AV} between the auditory and visual parameter estimates of the linear regression (w_{AV} = atan($ß_V$ / $ß_A$)). Thus, w_{AV} varied between pure visual (90°) and pure auditory (0°) influence (Fig. 4.4A-D). Confidence intervals of the mean w_{AV} were computed using a double bootstrap (Martin, 1990) (to account for the small number of participants) for circular measures. To refrain from making any parametric assumptions, we evaluated the main effects of visual reliability, task-relevance, spatial discrepancy and their interactions in the factorial design using permutation testing of a likelihood ratio test statistic (Anderson and Wu, 1995) for circular measures (Tab. 4.1). To account for the within-subject design, permutations were constrained to occur within each participant. For the main effects of visual reliability, task-relevance and discrepancy, w_{AV} values were permuted within the levels of the non-tested factors (5000 random permutations). For tests of the interactions, values were freely permuted across all conditions (Gonzalez and Manly, 1998) (5000 random permutations).

Further, we identified multisensory influences in primary sensory areas (i.e., auditory influence on V1 and visual influence on A1) for small spatial discrepancies by testing whether w_{AV} was smaller than 90° (in V1) or larger than 0° (in A1) using a one-sided permutation test. Specifically for small spatial discrepancies (≤ 6.6°) (Stein and Meredith, 1993), we randomly assigned (5000 times) the sign of the circular distance from the critical value in each participant (pooling over visual reliability x task-relevance). Unless otherwise stated, results are reported at $p < 0.05$.

To correlate the neural and behavioral audiovisual weight indices, we first computed the behavioral audiovisual weight index using an equivalent regression model as for the neural weight index but with participants' behavioral responses (instead of the decoded spatial location) being the dependent variable. Using a circular-circular correlation, we computed the correlation coefficient between neural and behavioral weight indices for each participant and condition in the two (visual reliability: high vs. low) x two (task-relevance: auditory vs. visual report) x two (spatial discrepancy: ≤ 6.6° vs. > 6.6°) factorial design (Fig. 4.4C). The correlation coefficients were averaged across participants after Fisher's z transformation.

Auditory and visual retinotopic localizer

Auditory and visual retinotopic localizers were used to define regions of interest along the auditory and visual processing hierarchies in a participant-specific fashion. In the auditory localizer, participants were presented with brief bursts of white noise at -10° or 10° visual angle (duration 500 ms, stimulus onset asynchrony 2 s). In a one-back task, participants indicated via a key press when the spatial location of the current trial was different from the previous trial. 20 s blocks of auditory conditions (i.e., 20 trials) alternated with 13 s fixation periods. The auditory locations were presented in a pseudorandomized fashion to optimize design efficiency. Similar to the main experiment, the auditory localizer sessions were modelled in an event-related fashion with the onset vectors of left and right auditory stimuli being entered into the design matrix after convolution with the hemodynamic response function and its first temporal derivative. Auditory responsive regions were defined as voxels in superior temporal and Heschl's gyrus showing significant activations for auditory stimulation relative to fixation ($p < 0.05$, family-wise error corrected). Within these regions, we defined primary auditory cortex (A1) based on cytoarchitectonic probability maps (Eickhoff et al., 2005) and referred to the remainder (i.e., planum temporale and posterior superior temporal gyrus) as higher order auditory cortex (hA, see fig. 4.3).

Standard phase-encoded retinotopic mapping (Sereno et al., 1995) was used to define visual regions of interest. Participants viewed a checkerboard background flickering at 7.5 Hz through a rotating wedge aperture of 70° width (polar angle mapping) or an expanding/contracting ring (eccentricity mapping). The periodicity of the apertures was 42 s. Visual responses were modelled by entering a sine and cosine convolved with the hemodynamic response function as regressors in a general linear model. The preferred polar angle was determined as the phase lag for each voxel which is the angle between the parameter estimates for the sine and the cosine. The preferred phase lags for each voxel were projected on the reconstructed, inflated cortical surface using Freesurfer 5.1.0 (Dale et al., 1999). Visual regions V1-V3, V3AB and IPS0-IPS4 were defined as phase reversal in angular retinotopic maps. IPS0-4 were defined as contiguous, approximately rectangular regions based on phase reversals along the anatomical IPS (Swisher et al., 2007). For the decoding analyses, the auditory and visual regions were combined from the left and right hemispheres.

Control analyses to account for eye movements as potential confounds

Eye recordings were calibrated with standard eccentricities between ±3° and ±10° to determine the deviation from the fixation cross. Fixation position was post-hoc offset corrected. Eye position data were automatically corrected for blinks and converted to radial velocity. For each trial, the post-stimulus mean horizontal eye position and the number of saccades (defined by a radial eye-velocity threshold of 15° s^{-1} for a minimum of 60 ms duration and radial amplitude larger than 1°) were quantified (0-875 ms after stimulus onset). We analyzed the eye movement indices including percent saccades, percent eye blinks and the post-stimulus mean horizontal eye position in three separate four (visual location) x four (auditory location) x two (visual reliability) x two (task-relevance) repeated measure ANOVAs. As only the analysis of mean horizontal eye position revealed a trend of task relevance (Supplemental tab. S4.1), we performed an additional control analysis where we included the post-stimulus mean horizontal eye position as a nuisance covariate into the regression model in addition to the true auditory and visual locations to predict the fMRI decoded locations (Supplemental fig. S4.1 and tab. S4.2).

4.4 Results

In a spatial ventriloquist paradigm, we scanned five participants with fMRI whilst they were presented with synchronous, yet spatially congruent or discrepant visual and auditory signals (Fig. 4.1A). Visual and auditory locations were independently sampled from four spatial locations along the azimuth (i.e., -10°, -3.3°, 3.3° or 10°) resulting in four possible levels of spatial discrepancy (0°, 6.6°, 13.3°, 20°). In addition, we manipulated the reliability of the visual signal (high vs. low). On each trial, participants either reported their auditory or visual perceived location. Thus, the ventriloquist paradigm factorially manipulated auditory location, visual location, visual reliability and task-relevance (Fig. 4.1B). Yet, to investigate information integration from the perspective of causal inference, we reorganized these conditions into a two (visual reliability: high vs. low) x two (task-relevance: auditory vs. visual report) x two (spatial discrepancy: ≤ 6.6° vs. > 6.6°) factorial design for the statistical analysis of the behavioral and fMRI data.

Behavioral results

The crossmodal bias [$CMB = (L_{response} – L_{Auditory}) / (L_{Visual} – L_{Auditory})$] quantifies the relative influence of the visual and auditory signals on the reported auditory and visual locations (Fig. 4.1C). We evaluated the crossmodal bias in a two (visual reliability: high vs. low) x two (task-relevance: auditory vs. visual report) x two (spatial discrepancy: ≤ 6.6° vs. > 6.6°) repeated measure ANOVA.

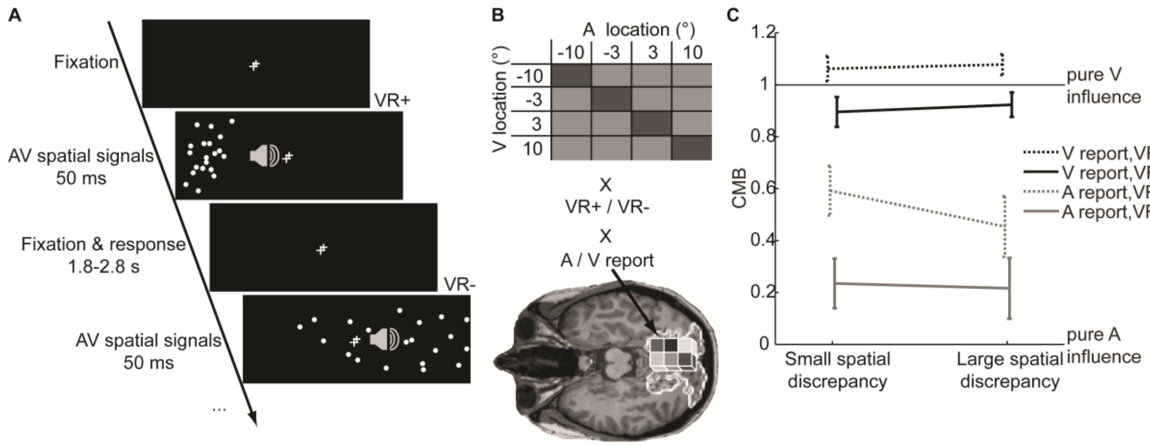


Figure 4.1. Example trial, experimental design and behavioral data. (A) In a ventriloquist paradigm, participants were presented with synchronous audiovisual (AV) signals originating from four possible locations along the azimuth. The visual signal was a cloud of white dots. The auditory signal was a brief burst of white noise. Participants localized either the auditory or the visual signal (n.b. for illustrational purposes the visual angles of the cloud have been scaled in a non-uniform fashion in this scheme). **(B)** The four-factorial experimental design manipulated (1) the location of the visual (V) signal (-10°, -3.3°, 3.3°, 10°) (2) the location of the auditory (A) signal (-10°, -3.3°, 3.3°, 10°), (3) the reliability of the visual signal (high (VR+) versus low (VR-) spread of the visual cloud) (4) task-relevance (auditory vs. visual report). Using fMRI, we measured activation patterns to audiovisual signals of all experimental conditions from voxels of regions along the auditory and visual spatial processing hierarchies. **(C)** Crossmodal bias (CMB; across participants mean ± SEM, N = 5) as a function of visual reliability, task-relevance and spatial discrepancy (small (≤ 6.6°) vs. large (> 6.6°)). CMB quantifies the relative influence of the auditory and the visual signal on the reported locations. If CMB equals one, the reported location is influenced purely by the visual signal. If CMB equals zero, the reported location is influenced purely by the auditory signal.

In line with the principle of reliability-weighted integration, the relative visual influence on the reported auditory and visual locations increased for high relative to low visual reliability as indicated by a significant main effect of visual reliability ($F_{1,4} = 27.329$, $p = 0.006$). Yet, even though the perceived (and reported) auditory location shifted towards the concurrent visual signal and vice versa (for small spatial discrepancies), the perceived auditory and visual locations differed for identical audiovisual stimulus combinations as indicated by a significant main effect of task-relevance on the crossmodal bias ($F_{1,4} = 41.372$, $p = 0.003$). This response profile suggests that audiovisual signals were not fused into one unified percept.

Critically, in line with Bayesian causal inference, this difference between crossmodal biases for auditory and visual report was significantly increased for large (> 6.6°) spatial discrepancies, when it is more likely that auditory and visual signals are caused by independent sources (i.e., a significant interaction between task-relevance and spatial discrepancy: $F_{1,4} = 40.232$, $p = 0.003$). Conversely, the modulatory effect of reliability on the crossmodal bias significantly decreased for large relative to small spatial discrepancies (i.e., interaction between reliability and spatial discrepancy: $F_{1,4} = 9.508$, $p = 0.037$).

Collectively, our behavioral results suggest that humans integrate audiovisual signals into spatial representations in line with the principles of Bayesian causal inference. For small spatial discrepancy, audiovisual signals are predominantly integrated weighted by the relative sensory reliabilities. For large spatial discrepancy, audiovisual integration is attenuated and the reported locations depend more strongly on the reported sensory modality.

Functional Imaging analysis: fMRI decoding strategy

To characterize how auditory and visual signals are integrated into spatial representations along the dorsal visual (Mishkin et al., 1983) and the auditory (Tian et al., 2001) spatial processing hierarchies, we combined functional magnetic resonance imaging (fMRI) with a multivariate pattern decoding approach. First, we defined the mapping between fMRI activation pattern and the spatial locations in the external world by training a support vector regression model selectively on the 16 types of trials that present auditory and visual signals at *congruent* locations along the azimuth (Fig. 4.2A). This trained support vector regression model was then used to decode the spatial locations from independent activation patterns of spatially *congruent and discrepant* trials (cf. fig. 4.1B).

Second, we quantified the influence of auditory and visual signals on the spatial locations decoded from fMRI activation patterns separately for each region of interest using a linear regression model (Fig. 4.2B). In this regression model, the fMRI decoded locations were predicted by 16 regressors modelling the true spatial locations separately for auditory and visual signals for each condition in our two (auditory vs. visual report) x two (high vs. low visual reliability) x two (large vs. small spatial discrepancy) factorial design. The parameter estimates quantified the influence of the true auditory and visual locations on the decoded spatial representations. For instance, a high visual (ß_V) and low auditory (ß_A) parameter estimate suggests that a region integrates auditory and visual signals with a stronger weight assigned to the visual signal. In a region integrating audiovisual signals in line with Bayesian causal inference, we would expect that the auditory and visual regression coefficients depend on visual reliability and task-relevance (i.e., auditory vs. visual report; Fig. 4.2C).

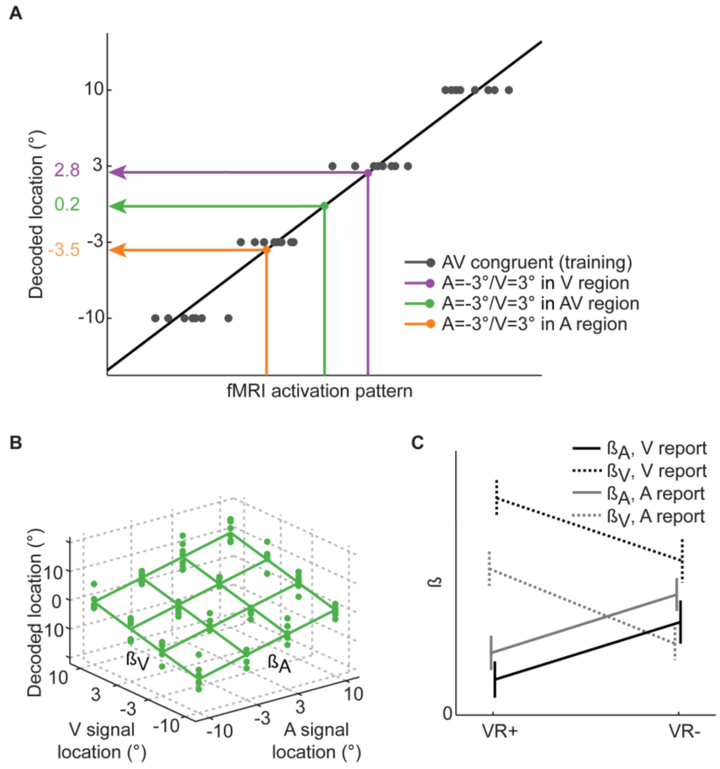


Figure 4.2. Multivariate pattern analysis to decode spatial representations from fMRI activation patterns. (A) Based on audiovisual congruent conditions alone, a linear support vector regression (SVR) model learnt the mapping from activation patterns to signal locations. This trained SVR model was used to decode the spatial representations from independent test activation patterns of both audiovisual incongruent and congruent trials. For instance, the activation pattern for audiovisual (A = -3° / V = 3°) incongruent trials should map approximately to 3° in a purely visual (V) region and to -3° in a purely auditory (A) region. In an audiovisual integration region, however, it maps to an intermediate value between -3° and 3° depending on the relative influences of auditory and visual signals on the activation patterns. **(B)** The auditory and visual influences were quantified in a linear regression that used the true auditory and visual locations (separately for the conditions of visual reliability x task-relevance x spatial discrepancy) to predict the fMRI decoded signal locations. The auditory (ß$_A$) and visual (ß$_V$) regression coefficients index the influence of auditory and/or visual spatial signals on a region's spatial representation. **(C)** In a region that only partially integrates audiovisual signals weighted by reliability, we predicted that a decrease in visual reliability (high (VR+) vs. small (VR-)) reduces the visual (i.e., ß$_V$) and concurrently increases the auditory influence (i.e., ß$_A$) on its spatial representations (pooling over spatial discrepancy). Further, we predicted that auditory report increases auditory influence, while visual report increases visual influence.

Third, we obtained a summary index for the relative audiovisual weights by computing w$_{AV}$ as the inverse tangent of the visual and auditory parameter estimates from the regression model (= atan(ß$_V$ / ß$_A$)). This relative weight index ranges from pure visual (90°) to pure auditory (0°) influence (Fig. 4.4A-D). We performed the statistics on the audiovisual neural weight index using a two (auditory vs. visual report) x two (high vs. low visual reliability) x two (large vs. small spatial discrepancy) factorial design (see tab. 4.1 for results of the three-way analysis based on circular statistics (Anderson and Wu, 1995)). As the three-way interaction was not significant, we present the parameter estimates from the regression model and the relative audiovisual weights separately as a function of visual reliability, task-relevance and spatial discrepancy (Fig. 4.3 and 4.4A-C) and as a function of both task-relevance and spatial discrepancy (Fig. 4.4D).

Table 4.1. Statistical significance of main and interaction effects of the factors visual reliability (VR), task-relevance (TR) and discrepancy (Discr) for the audiovisual weight index w_{AV}.

	VR	TR	Discr	VR X TR	VR X Discr	TR X Discr	VR X TR X Discr
	p	p	p	p	p	p	p
V1	0.96	0.18	**<0.001**	0.49	0.07	0.88	0.65
V2	**0.02**	0.62	**0.008**	0.47	0.29	0.25	0.23
V3	0.64	0.20	0.24	0.34	0.29	0.67	0.66
V3AB	0.47	0.06	0.07	0.61	**0.01**	0.96	0.54
IPS0-2	**0.009**	**0.003**	0.79	0.08	0.13	0.94	0.94
IPS3-4	**<0.001**	**<0.001**	0.50	0.49	0.57	**0.05**	0.61
hA	0.41	**0.001**	0.55	0.44	0.89	0.44	0.75
A1	0.09	0.09	0.10	0.57	0.96	0.80	0.28

Note: p values were derived from permutation tests using a circular log likelihood ratio statistic. N = 5. P values in bold indicate significant values.

Audiovisual influences in primary sensory areas

As accumulating evidence suggests that multisensory integration may start already at the primary, putatively unisensory level (Foxe et al., 2000; Bonath et al., 2007; Kayser et al., 2007; Lakatos et al., 2007; Noesselt et al., 2007; Lewis and Noppeney, 2010; Werner and Noppeney, 2010), we first investigated whether audiovisual influences can be identified in primary auditory and visual areas. For small audiovisual spatial discrepancies (≤ 6.6°), we expected that the spatial location decoded from primary auditory areas shifts towards the true visual location and that the location decoded from primary visual areas shifts towards the true auditory location. Indeed, in primary visual area (V1) the relative audiovisual weight index was significantly smaller than 90° ($p = 0.022$, one-sided permutation test) indicating that the decoded spatial locations were biased towards the true auditory location. Conversely, in primary auditory cortex (A1) the audiovisual weight index was larger than 0° ($p = 0.008$, one-sided permutation test) indicating a significant bias towards the true visual location. While these results demonstrate that signals from the non-preferred sensory modality influence neural representations at the primary cortical level, the crossmodal influences were relatively small. Thus, in primary sensory areas, the decoded location was predominantly determined by the true location of the preferred sensory signal (Fig. 4.4B).

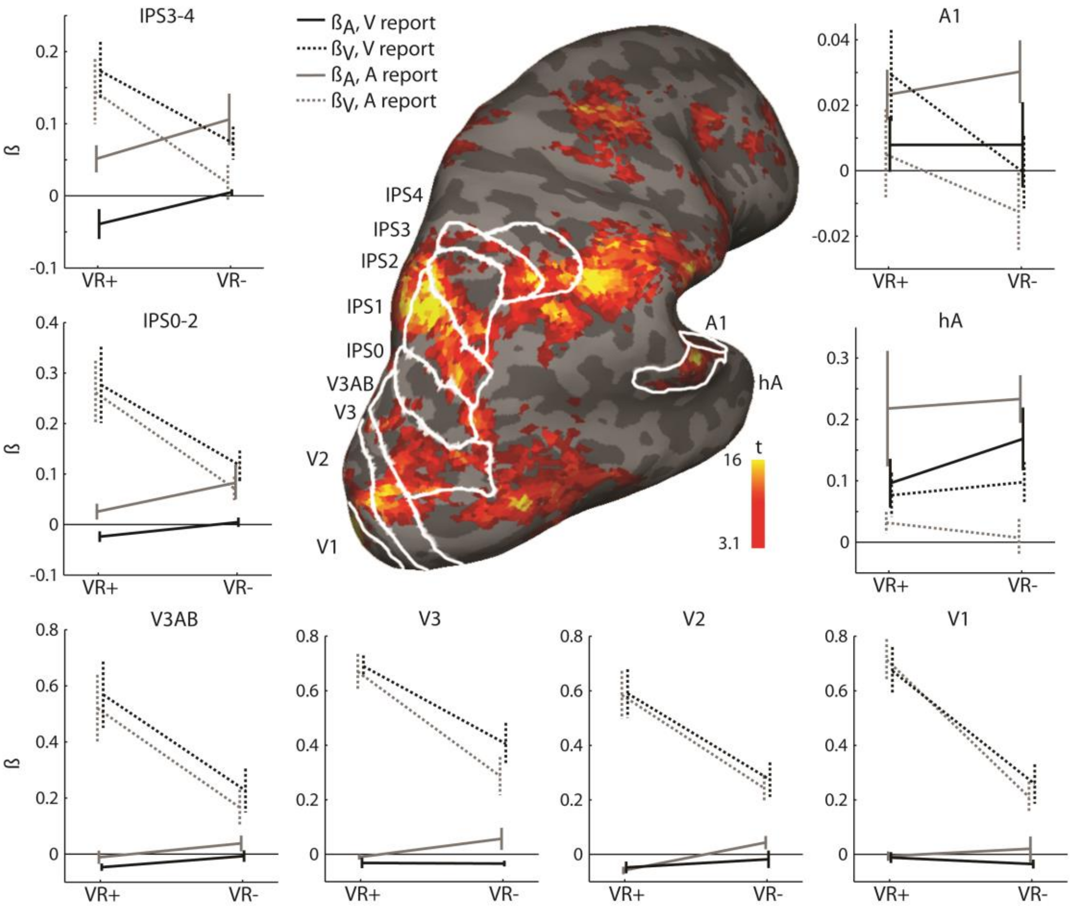


Figure 4.3. The influence of auditory and visual signals on the spatial representations decoded from regions along the auditory and visual spatial processing hierarchies. The influence – as indexed by the auditory ($ß_A$) and visual ($ß_V$) regression coefficients (across participants mean ± SEM; N = 5) - are shown as a function of visual reliability (high (VR+) vs. small (VR-)) and task-relevance (auditory (A) vs. visual (V) report) (pooled over spatial discrepancy). In the center of the figure, activations for all audiovisual stimuli > baseline for a representative participant is overlaid on the participant's inflated cortical surface (thresholded at $p < 0.001$, uncorrected for illustrational purposes). The functionally defined regions of interest are demarcated in white.

Effect of sensory reliability on audiovisual integration

Behaviorally, humans integrate signals that are close in space, time and structure weighted by their relative reliabilities (Ernst and Banks, 2002; Alais and Burr, 2004). Yet, low-level visual areas encompassing V1, V2, V3 and V3A/B showed an effect of visual reliability predominantly on the visual parameter estimate (see fig. 4.3, bottom). By contrast, only higher parietal cortices (IPS0 - IPS4) were governed by the classical reliability-driven reweighting where a decrease in visual reliability induced a concurrent reduction in visual and an increase in auditory weight (see figure 4.3, top left). Indeed, these impressions were confirmed in the circular statistics of the relative audiovisual weight index that identified significant effects of reliability primarily in IPS0-2 and IPS3-4. An additional effect of reliability was observed in V2.

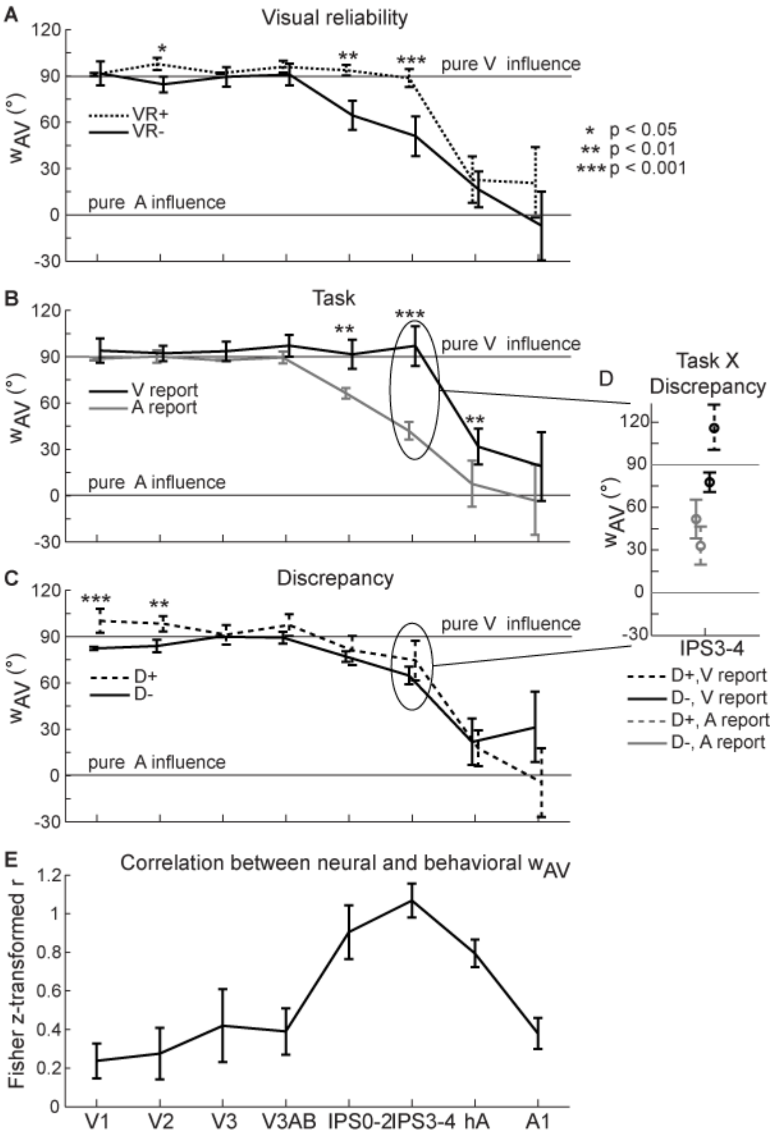


Figure 4.4. Audiovisual weight index as a function of visual reliability, task-relevance and discrepancy and its correlation with the corresponding behavioral weight index in the regions of interest. Audiovisual weight index w_{AV} (across participants circular mean and double-bootstrapped 68% confidence interval, N = 5) was computed as the angle between the auditory and visual regression coefficients ($atan(ß_V/ß_A)$). For a purely visual region, w_{AV} is 90°. For a purely auditory region, it is 0°. Asterisks indicate the statistical significance of effects on w_{AV} derived from a circular log likelihood ratio statistic. (A) Audiovisual weight index w_{AV} as a function of visual reliability (high (VR+) vs. small (VR-)). (B) Audiovisual weight index w_{AV} as a function of task-relevance (auditory (A) vs. visual (V) report). (C) Audiovisual weight index w_{AV} as a function of audiovisual spatial discrepancy (small (≤ 6.6°; D-) vs. large (> 6.6°; D+)). (D) Audiovisual weight index w_{AV} in IPS3-4 as a function of task-relevance and discrepancy. (E) Circular-circular correlation (across participants mean after Fisher z transformation ± SEM, N = 5) between the neural weight index w_{AV} and the equivalent behavioral weight index in the regions of interest.

Effect of task-relevance on audiovisual integration and its interaction with spatial discrepancy

Next, we asked where auditory and visual signals are integrated into spatial representations depending on whether the visual or the auditory signals were task-relevant and needed to be reported as observed in human behavior. While we found a significant main effect of task-relevance (i.e., visual vs. auditory report) on the audiovisual weight index already in higher-order auditory areas (hA) encompassing the belt and the planum temporale, the effect emerged predominantly in higher-order association areas such as IPS0-4 (cf. tab. 4.1). In all these areas, the visual signal exerted a stronger influence on the decoded location during visual report and the auditory signal on the decoded location during auditory report (Fig. 4.3 top left; Fig. 4.4B).

The difference in spatial representations for auditory and visual report already suggests that the brain integrates audiovisual signals only partially. More formally, it combines a fused multisensory estimate with a unimodal visual (or auditory) estimate for visual (or auditory) report. From the perspective of causal inference, we expected the influence of the unisensory estimate on the spatial representations to be stronger for large spatial discrepancies, when it is more likely that auditory and visual signals are generated by independent sources. Indeed, IPS3-4 showed a significant interaction between task-relevance and spatial discrepancy (cf. tab. 4.1). Thus, as expected under Bayesian causal inference, the spatial representations decoded from IPS3-4 showed a greater difference for visual versus auditory report for large (Fig. 4.4D, dashed lines) relative to small spatial discrepancies (Fig. 4.4D, solid lines). These results indicate that IPS3-4 integrates the audiovisual signals by taking into account the probabilities of the potential causal structures in the environment.

Effect of spatial discrepancy on audiovisual integration

In contrast to the interaction between task-relevance and spatial discrepancy that was found primarily in IPS areas, we observed a main effect of spatial discrepancy in V1 and V2 (Fig. 4.4C). Small spatial discrepancies increased the auditory ‘attractive’ influence on the spatial representations decoded from visual areas (c.f. solid lines are below dotted lines in V1, V2 in Fig. 4.4C). Conversely, small spatial discrepancy increased the visual ‘attractive’ influence on spatial representations decoded from auditory areas (solid lines are above dotted lines in A1 in fig. 4.4C, $p < 0.05$ for unidirectional hypothesis). These results suggest that integration in low-level sensory areas depends on auditory and visual signals co-occurring within a spatial window (Stein and Meredith, 1993).

Relation of neural and behavioral weight indices of audiovisual spatial integration

Finally, we asked how and where in the cortical hierarchies the neural weights were related with the behavioral weights. For this, we computed the behavioral weights based on participants’ behavioral localization reports using the same regression approach that we employed for the fMRI decoded spatial locations. Next, we computed the correlation between the neural and behavioral weight indices for each of the regions of interest. The correlation coefficient increased along the visual processing hierarchy culminating in IPS3-4 (Fig. 4.4E). Likewise, in the auditory system, the correlation between neural and behavioral weights was enhanced in higher-order auditory areas relative to primary auditory cortex. Hence, IPS3-4 integrates auditory and visual signals into spatial representations that are critical to guide behavioral performance such as spatial orienting.

Controlling for eye movements as potential confounds

To address potential concerns that our decoding results may be confounded by eye movements, we performed a series of control analyses. First, we evaluated participants' eye movements based on eye tracking data recorded concurrently during fMRI acquisition. Fixation was well maintained throughout the experiment with post-stimulus saccades detected in only 2.293 ± 1.043 % (mean ± SEM) of the trials. Moreover, 4 (visual location) x 4 (auditory location) x 2 (visual reliability) x 2 (task-relevance) repeated measure ANOVAs performed separately for (i) % saccades or (ii) % eye blinks revealed no significant main effects or interactions. The repeated measure ANOVA on post-stimulus mean horizontal eye position (0-875 ms post-stimulus onset) revealed only trends for the main effect of task-relevance and visual local positions (Supplemental tab. S4.1).

As a further control analysis, we therefore re-performed the linear regression analyses (with fMRI decoded spatial location as dependent variable) and included post-stimulus mean horizontal eye position as a nuisance covariate in addition to the true auditory and visual locations to predict the fMRI decoded locations. This analysis basically replicated our initial results (Supplemental fig. S4.1 and tab. S4.2).

4.5 Discussion

This study combined psychophysics and fMRI to investigate how the human brain integrates and segregates sensory signals depending on the causal structure of the multisensory world.

At the behavioral level, our results show that participants integrate signals into spatial representations in line with the principles of Bayesian causal inference: They integrate audiovisual signals weighted by their reliability, when signals are close in space. They mostly segregate information and report the location predominantly of the task-relevant signal, when signals are spatially discrepant and hence unlikely to arise from a common source.

Combining fMRI and multivariate pattern decoding we characterized how the brain forms audiovisual spatial representations along the visual (Mishkin et al., 1983) and auditory (Tian et al., 2001) spatial processing hierarchies. Accumulating evidence has demonstrated that multisensory integration is not limited to association cortices (Beauchamp et al., 2004; Hein et al., 2007; Sadaghiani et al., 2009; Lewis and Noppeney, 2010; Werner and Noppeney, 2010), but emerges already at the primary, putatively unisensory, level (Foxe et al., 2000; Bonath et al., 2007; Kayser et al., 2007; Lakatos et al., 2007; Noesselt et al., 2007; Lewis and Noppeney, 2010; Werner and Noppeney, 2010) via thalamocortical mechanisms (Lakatos et al., 2007), direct connectivity between sensory areas (Falchier et al., 2002) or top-down influences from higher-order association cortices

(Macaluso and Driver, 2005). Likewise, we observed bidirectional audiovisual influences already in primary sensory areas: The visual location influenced the spatial representations in primary auditory cortex and vice versa. In line with the well-established spatial principle of multisensory integration (Stein and Meredith, 1993), these audiovisual influences emerged primarily when auditory and visual signals were close in space. Yet, even in the case of small spatial discrepancy, the audiovisual influences in primary sensory areas were still relatively small when compared to parietal cortices. These findings dovetail nicely with previous neurophysiological studies showing about 15% 'multisensory' neurons in primary sensory areas (Bizley et al., 2007) but more than 50% in classical association areas such as intraparietal or superior temporal sulci (Dahl et al., 2009).

Critically, however, our study showed that multisensory interactions not only increased progressively along the cortical hierarchy, but also changed their computational operations from primary visual to higher-order parietal areas. In primary visual areas, the visual reliability predominantly affected the visual influence on the decoded spatial locations leaving the auditory influence mostly unchanged. In other words, low level sensory areas did not yet combine sensory inputs according to reliability-driven reweighting (see fig. 4.3).

By contrast, higher-order parietal areas (IPS0-4) integrated auditory and visual signals weighted by their reliability, such that a decrease in visual reliability reduced the influence of the visual signals and concurrently amplified the influence of the auditory signals on the decoded spatial location. Yet, even parietal areas did not integrate sensory signals into one unified or 'amodal' spatial representation as expected under traditional forced fusion models (Ernst and Banks, 2002; Alais and Burr, 2004). Instead, IPS0-4 integrated auditory and visual spatial inputs only partially, so that the decoded spatial estimates differed for identical audiovisual signals depending on the reported sensory modality (cf. fig. 4.4B). These results suggest that higher-order parietal cortices integrate audiovisual signals weighted by both bottom-up sensory reliability and top-down task-relevance (i.e., visual vs. auditory report) (Gottlieb et al., 1998).

Critically, in line with Bayesian causal inference, spatial discrepancy increased the difference between the spatial estimates decoded from IPS3-4 for auditory and visual report: When auditory and visual signals were spatially proximate and likely to provide information about the same event, IPS3-4 integrated them weighted by their reliability into spatial estimates that converged for auditory and visual report. Yet, when audiovisual signals were spatially discrepant and hence likely to emanate from independent events, IPS3-4 processed sensory information in a predominantly segregated fashion. In sum, Bayesian causal inference provides IPS3-4 with a computational strategy to arbitrate flexibly between information integration and segregation depending on the probabilities of the underlying causal structures.

Collectively, our results reveal spatial discrepancy as a critical cue informing the brain whether or not sensory signals are generated by independent sources and should hence be segregated. Importantly, however, spatial discrepancy induces distinct types of information segregation at low and high levels of the cortical hierarchy. In primary sensory areas, spatial discrepancy controls the contribution of the non-preferred signals to the spatial representations irrespective of the task-context. For instance, large spatial discrepancies enable the primary visual cortex to form a visual spatial estimate largely unaffected (or even repulsed) by concurrent auditory input (and vice versa for auditory areas).

By contrast, in IPS3-4 spatial discrepancy controls the contribution of signals from the task-irrelevant sensory modality to the neural spatial representations. Exploiting information from all sensory signals depending on whether they arise from common or independent sources allows IPS3-4 to compute the most precise spatial estimates of the task-relevant target position. Indeed, the behavioral relevance of the IPS3-4 spatial estimates is further indicated by the correlation between the neural and behavioral weights, which progressively increases along the auditory and visual processing hierarchies. It culminates in IPS3-4, which plays a key role in guiding behavioral responses for spatial tasks (Macaluso et al., 2003).

Our results thus suggest that parietal cortices integrate sensory signals into multisensory spatial priority maps (Busse et al., 2005; Talsma et al., 2010) where the relevance of a spatial location is defined jointly by signals from multiple sensory modalities dependent on their relative sensory reliability, their importance for a particular task and the causal structure of the environment. Multisensory priority maps go functionally beyond traditional unisensory spatial priority maps (Gottlieb et al., 1998; Bisley and Goldberg, 2010), as they enable spatial orienting and effective interactions with our complex multisensory environment (Macaluso et al., 2003). Via back-projections these IPS spatial priority maps may also mediate audiovisual influences on spatial representations in low-level sensory areas (Macaluso and Driver, 2005; Driver and Noesselt, 2008) thereby making prioritized locations available to large parts of neocortex (Ghazanfar and Schroeder, 2006).

In conclusion, our results demonstrate distinct computational operations in low-level sensory and higher-order association areas. In low-level sensory areas, multisensory influences were small, not yet governed by reliability-driven reweighting and less susceptible to top-down influences. By contrast, IPS3-4 partially integrated sensory signals weighted by their bottom-up reliability and top-down task-relevance into spatial representations that take into account the probabilities of the environmental causal structures. Thus, IPS3-4 integrates sensory signals into multisensory spatial priority maps in accordance with the principles of Bayesian causal inference.

4.6 Acknowledgments

This study was funded by the Max Planck Society and the European Research Council. We thank C. Frith and U. Beierholm for very helpful comments on a previous draft of this manuscript and P. Ehses for advice on the fMRI sequence. We thank S. Schwarzkopf for providing stimulus scripts for retinotopic mapping (http://sampendu.wordpress.com/retinotopy-tutorial/).

4.7 References

Alais D, Burr D (2004) The ventriloquist effect results from near-optimal bimodal integration. Curr Biol 14:257-262.

Algazi VR, Duda RO, Thompson DM, Avendano C (2001) The cipic hrtf database. In: Applications of Signal Processing to Audio and Acoustics, 2001 IEEE Workshop on the, pp 99-102: IEEE.

Anderson CM, Wu CJ (1995) Measuring location effects from factorial experiments with a directional response. International Statistical Review/Revue Internationale de Statistique:345-363.

Beauchamp MS, Argall BD, Bodurka J, Duyn JH, Martin A (2004) Unraveling multisensory integration: patchy organization within human STS multisensory cortex. Nat Neurosci 7:1190-1192.

Bertelson P, Radeau M (1981) Cross-modal bias and perceptual fusion with auditory-visual spatial discordance. Attention, Perception, & Psychophysics 29:578-584.

Bisley JW, Goldberg ME (2010) Attention, intention, and priority in the parietal lobe. Annu Rev Neurosci 33:1-21.

Bizley JK, Nodal FR, Bajo VM, Nelken I, King AJ (2007) Physiological and anatomical evidence for multisensory interactions in auditory cortex. Cereb Cortex 17:2172-2189.

Bonath B, Noesselt T, Martinez A, Mishra J, Schwiecker K, Heinze HJ, Hillyard SA (2007) Neural basis of the ventriloquist illusion. Curr Biol 17:1697-1703.

Brainard DH (1997) The psychophysics toolbox. Spatial vision 10:433-436.

Busse L, Roberts KC, Crist RE, Weissman DH, Woldorff MG (2005) The spread of attention across modalities and space in a multisensory object. Proc Natl Acad Sci U S A 102:18751-18756.

Chang CC, Lin CJ (2011) LIBSVM: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology (TIST) 2:27.

Dahl CD, Logothetis NK, Kayser C (2009) Spatial organization of multisensory responses in temporal association cortex. J Neurosci 29:11924-11932.

Dale AM, Fischl B, Sereno MI (1999) Cortical surface-based analysis. I. Segmentation and surface reconstruction. Neuroimage 9:179-194.

Driver J, Noesselt T (2008) Multisensory interplay reveals crossmodal influences on 'sensory-specific' brain regions, neural responses, and judgments. Neuron 57:11-23.

Eickhoff SB, Stephan KE, Mohlberg H, Grefkes C, Fink GR, Amunts K, Zilles K (2005) A new SPM toolbox for combining probabilistic cytoarchitectonic maps and functional imaging data. Neuroimage 25:1325-1335.

Ernst MO, Banks MS (2002) Humans integrate visual and haptic information in a statistically optimal fashion. Nature 415:429-433.

Falchier A, Clavagnier S, Barone P, Kennedy H (2002) Anatomical evidence of multimodal integration in primate striate cortex. J Neurosci 22:5749-5759.

Fetsch CR, Deangelis GC, Angelaki DE (2013) Bridging the gap between theories of sensory cue integration and the physiology of multisensory neurons. Nat Rev Neurosci 14:429-442.

Fetsch CR, Pouget A, DeAngelis GC, Angelaki DE (2012) Neural correlates of reliability-based cue weighting during multisensory integration. Nat Neurosci 15:146-154.

Foxe JJ, Morocz IA, Murray MM, Higgins BA, Javitt DC, Schroeder CE (2000) Multisensory auditory-somatosensory interactions in early cortical processing revealed by high-density electrical mapping. Brain Res Cogn Brain Res 10:77-83.

Friston KJ, Holmes AP, Worsley KJ, Poline JP, Frith CD, Frackowiak RSJ (1994) Statistical parametric maps in functional imaging: a general linear approach. Human brain mapping 2:189-210.

Ghazanfar AA, Schroeder CE (2006) Is neocortex essentially multisensory? Trends Cogn Sci 10:278-285.

Gonzalez L, Manly B (1998) Analysis of variance by randomization with small data sets. Environmetrics 9:53-65.

Gottlieb JP, Kusunoki M, Goldberg ME (1998) The representation of visual salience in monkey parietal cortex. Nature 391:481-484.

Hein G, Doehrmann O, Muller NG, Kaiser J, Muckli L, Naumer MJ (2007) Object familiarity and semantic congruency modulate responses in cortical audiovisual integration areas. J Neurosci 27:7881-7887.

Kayser C, Petkov CI, Augath M, Logothetis NK (2007) Functional imaging reveals visual modulation of specific fields in auditory cortex. J Neurosci 27:1824-1835.

Kording KP, Beierholm U, Ma WJ, Quartz S, Tenenbaum JB, Shams L (2007) Causal inference in multisensory perception. PLoS One 2:e943.

Lakatos P, Chen CM, O'Connell MN, Mills A, Schroeder CE (2007) Neuronal oscillations and multisensory interaction in primary auditory cortex. Neuron 53:279-292.

Lewis R, Noppeney U (2010) Audiovisual synchrony improves motion discrimination via enhanced connectivity between early visual and auditory areas. J Neurosci 30:12329-12339.

Liu T, Hospadaruk L, Zhu DC, Gardner JL (2011) Feature-specific attentional priority signals in human cortex. J Neurosci 31:4484-4495.

Ma WJ, Beck JM, Latham PE, Pouget A (2006) Bayesian inference with probabilistic population codes. Nat Neurosci 9:1432-1438.

Macaluso E, Driver J (2005) Multisensory spatial interactions: a window onto functional integration in the human brain. Trends Neurosci 28:264-271.

Macaluso E, Driver J, Frith CD (2003) Multimodal spatial representations engaged in human parietal cortex during both saccadic and manual spatial orienting. Curr Biol 13:990-999.

Martin MA (1990) On the double bootstrap. In: Computing Science and Statistics: Interface'90. Proceedings of the 22nd Symposium on the Interface, pp 73-78.

Mishkin M, Ungerleider LG, Macko KA (1983) Object vision and spatial vision: two cortical pathways. Trends in neurosciences 6:414-417.

Morgan ML, Deangelis GC, Angelaki DE (2008) Multisensory integration in macaque visual cortex depends on cue reliability. Neuron 59:662-673.

Noesselt T, Rieger JW, Schoenfeld MA, Kanowski M, Hinrichs H, Heinze HJ, Driver J (2007) Audiovisual temporal correspondence modulates human multisensory superior temporal sulcus plus primary sensory cortices. J Neurosci 27:11431-11441.

Parise CV, Spence C, Ernst MO (2012) When correlation implies causation in multisensory integration. Curr Biol 22:46-49.

Sadaghiani S, Maier JX, Noppeney U (2009) Natural, metaphoric, and linguistic auditory direction signals have distinct influences on visual motion processing. J Neurosci 29:6490-6499.

Sereno MI, Dale AM, Reppas JB, Kwong KK, Belliveau JW, Brady TJ, Rosen BR, Tootell RB (1995) Borders of multiple visual areas in humans revealed by functional magnetic resonance imaging. Science 268:889-893.

Shams L, Beierholm UR (2010) Causal inference in perception. Trends Cogn Sci 14:425-432.

Stein BE, Meredith MA (1993) The merging of the senses. Cambridge, MA: The MIT Press.

Swisher JD, Halko MA, Merabet LB, McMains SA, Somers DC (2007) Visual topography of human intraparietal sulcus. J Neurosci 27:5326-5337.

Talsma D, Senkowski D, Soto-Faraco S, Woldorff MG (2010) The multifaceted interplay between attention and multisensory integration. Trends Cogn Sci 14:400-410.

Tian B, Reser D, Durham A, Kustov A, Rauschecker JP (2001) Functional specialization in rhesus monkey auditory cortex. Science 292:290-293.

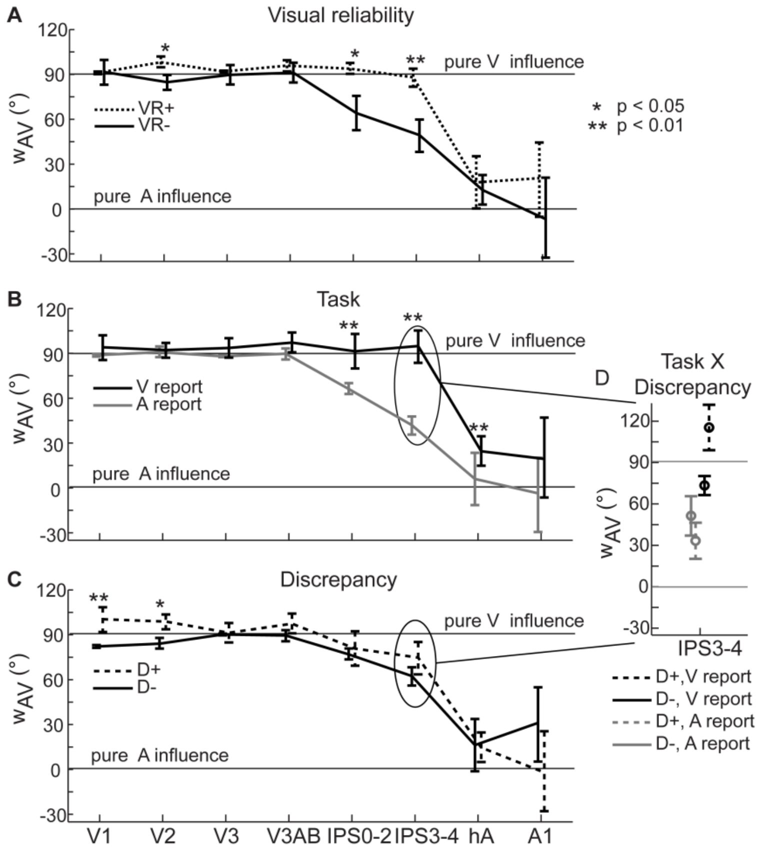
Wallace MT, Roberson GE, Hairston WD, Stein BE, Vaughan JW, Schirillo JA (2004) Unifying multisensory signals across time and space. Exp Brain Res 158:252-258.

Welch RB, Warren DH (1980) Immediate perceptual response to intersensory discrepancy. Psychol Bull 88:638-667.

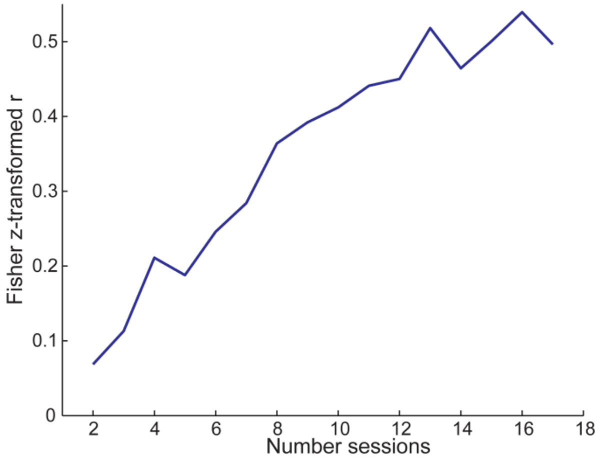
Werner S, Noppeney U (2010) Distinct functional contributions of primary sensory and association areas to audiovisual integration in object categorization. J Neurosci 30:2662-2675.

Wozny DR, Beierholm UR, Shams L (2010) Probability matching as a computational strategy used in perception. PLoS Comput Biol 6.

4.8 Supplemental results



Supplemental figure S4.1. Audiovisual weight index (after controlling for eye movements) in the regions of interest. Audiovisual weight index w_{AV} (across participants circular mean and double-bootstrapped 68% confidence interval, N = 5) was computed as the angle between the auditory and visual regression coefficients ($atan(ß_{V}/ß_{A})$). In order to control for horizontal eye movements, we included the post-stimulus mean horizontal eye position as a nuisance covariate in addition to the true auditory and visual locations to predict the fMRI decoded locations. For a purely visual region, w_{AV} is 90°. For a purely auditory region, it is 0°. Asterisks indicate the statistical significance of effects on w_{AV} derived from a circular log likelihood ratio statistic. **(A)** Audiovisual weight index w_{AV} as a function of visual reliability (high (VR+) vs. small (VR-)). **(B)** Audiovisual weight index w_{AV} as a function of task-relevance (auditory (A) vs. visual (V) report). **(C)** Audiovisual weight index w_{AV} as a function of audiovisual spatial discrepancy (small (≤ 6.6°; D-) vs. large (> 6.6°; D+)). **(D)** Audiovisual weight index w_{AV} in IPS3-4 as a function of task-relevance and discrepancy.



Supplemental figure S4.2. Decoding performance of the linear support vector regression model as a function of the number of included scanning sessions for a single participant from a prior pilot study. In a spatial ventriloquist paradigm, the pilot participant completed 33 sessions including 17 sessions for the auditory localization task. We computed the correlation coefficient between the true and decoded signal location in audiovisual congruent conditions as an index of decoding performance across an increasing number of sessions (i.e., from 2 sessions to 17 sessions). To obtain a more reliable estimate of the decoding performance, we computed the correlation coefficient by sampling n sessions without replacement up to 25 times for each number n of sessions. We computed and presented the correlation coefficient (after Fisher z transformation) averaged across these samples as a function of the number n of sessions.

Supplemental table S4.1. Main and interaction effects of the factors visual signal location (L_V), auditory signal location (L_A), visual reliability (VR), and task- relevance (TR) on post-stimulus eye-movement indices in repeated measure ANOVAs.

	Percent saccades				Horizontal eye position				Percent blinks			
	F	df1	df2	p	F	df1	df2	p	F	df1	df2	p
TR	3.82	1	4	0.12	7.73	1	4	0.05	0.04	1	4	0.86
VR	0.10	1	4	0.77	1.82	1	4	0.25	0.46	1	4	0.53
L_V	1.86	1.1	4.4	0.24	5.38	1.0	4.2	0.08	1.56	1.5	6.1	0.28
L_A	1.06	1.2	4.6	0.37	3.25	1.1	4.3	0.14	0.88	1.5	6.1	0.43
TR X VR	0.02	1	4	0.89	1.38	1	4	0.31	3.29	1	4	0.14
TR X L_V	1.70	1.0	4.2	0.26	5.84	1.0	4.2	0.07	0.79	1.9	7.4	0.48
TR X L_A	1.12	1.7	6.6	0.37	1.69	1.0	4.2	0.26	0.57	1.4	5.7	0.54
VR X L_V	1.02	2.0	7.8	0.40	3.08	1.6	6.3	0.12	3.26	2.0	7.9	0.09
VR X L_A	1.56	1.3	5.3	0.28	1.10	1.1	4.3	0.36	0.75	1.7	6.6	0.49
L_V X L_A	0.79	2.3	9.1	0.50	1.92	2.0	8.2	0.21	1.79	2.5	10.0	0.22
TR X VR X L_V	0.37	1.6	6.3	0.66	0.27	1.4	5.5	0.69	0.11	2.3	9.4	0.92
TR X VR X L_A	0.13	1.7	6.6	0.84	3.06	1.4	5.7	0.13	1.01	1.5	5.9	0.39
TR X L_V X L_A	0.97	2.3	9.3	0.43	0.62	2.4	9.8	0.59	1.63	2.6	10.3	0.24
VR X L_V x L_A	1.08	2.2	8.6	0.39	1.28	2.2	8.6	0.33	1.17	2.2	8.7	0.36
TR X VR X L_V X L_A	1.63	2.4	9.7	0.25	0.82	2.1	8.2	0.48	1.49	2.2	8.6	0.28

Note: p values are Greenhouse-Geisser corrected. N = 5.

Supplemental table S4.2. Statistical significance of main and interaction effects of the factors visual reliability (VR), task- relevance (TR) and discrepancy (Discr) for the audiovisual weight index w_{AV} when effects of the post-stimulus mean horizontal position of the eyes were controlled.

	VR	TR	Discr	VR X TR	VR X Discr	TR X Discr	VR X TR X Discr
	p	p	p	p	p	p	p
V1	0.96	0.18	**0.002**	0.51	**0.05**	0.85	0.65
V2	**0.03**	0.70	**0.024**	0.49	0.27	0.26	0.26
V3	0.64	0.24	0.71	0.38	0.32	0.78	0.67
V3AB	0.44	0.09	0.18	0.61	**0.004**	0.96	0.60
IPS0-2	**0.010**	**0.005**	0.93	0.08	0.11	0.90	0.96
IPS3-4	**0.001**	**0.001**	0.42	0.58	0.47	**0.05**	0.53
hA	0.54	**0.004**	0.83	0.32	0.87	0.36	0.68
A1	0.11	0.12	0.13	0.56	0.99	0.81	0.30

Note: p values were derived from permutation tests using a circular log likelihood ratio statistic. N = 5. P values in bold indicate significant values.

5 Suboptimal reliability-weighted integration of audiovisual spatial signals in parietal cortex

5.1 Abstract

To estimate an uncertain physical quantity, for example an object's location, the optimal strategy is to weight the quantity's noisy signals proportional to their relative reliability. Even though psychophysical studies demonstrate that human observers apply this principle when integrating unisensory, multisensory and motor signals, evidence for how the brain accomplishes this feat remains scarce. Combining psychophysics and multivariate fMRI decoding in a spatial ventriloquist paradigm, we characterized the computational operations underlying reliability-weighted audiovisual integration at several cortical levels along the auditory and visual processing hierarchy. The neural sensory weights were estimated by fitting 'neurometric' functions to the spatial locations decoded from regional fMRI activation patterns and compared to 'optimal' predicted sensory weights. Our results demonstrate that selectively the intraparietal sulcus forms spatial representations by integrating auditory and visual signals weighted by their relative reliability in a suboptimal fashion. Additionally, visual signals attained larger weights if they were selectively focused compared to a focus on auditory signals. By contrast, low-level auditory and visual regions encoded mainly the spatial signal of their preferred sensory modality, with only a small influence of the non-preferred modality. Together, the results demonstrate that higher-order multisensory regions perform probabilistic computations such as reliability-weighting, even though the computations might involve more complex processes like causal inference.

5.2 Introduction

In our natural environment our brain is confronted with noisy signals that provide uncertain information about the world. To construct the most likely and accurate representation of the environment, the brain is challenged to integrate signals from different senses if they pertain to a common object or event. Numerous psychophysics studies have demonstrated that human observers combine signals within and across the senses weighted in proportion to their reliability (i.e., the inverse of the signals' uncertainty)(Jacobs, 1999; Ernst and Banks, 2002; van Beers et al., 2002; Battaglia et al., 2003; Knill and Saunders, 2003; Alais and Burr, 2004; Hillis et al., 2004; Saunders and Knill, 2004; Rosas et al., 2005). In other words, greater weight is given to sensory signals that are more reliable. This reliability-weighted integration of sensory signals is statistically optimal in that it yields the most precise unbiased perceptual estimate (i.e., the maximum likelihood estimate, MLE). Thus, reliability-weighted integration is a fundamental mechanism that enables the brain to generate a more reliable representation of the world. Thereby, the weighting scheme increases performance accuracy on many tasks such as perception of depth (Jacobs, 1999), shape discrimination (Ernst and Banks, 2002) and spatial localization (Alais and Burr, 2004) as indicated by the ventriloquist illusion. Thus, in spatial ventriloquism the perceived sound location is shifted towards the visual stimulus and vice versa dependent on the relative reliabilities (Alais and Burr, 2004).

Despite the vast body of behavioral evidence showing near-optimal integration in humans, the neural basis of reliability-weighted integration has remained unexplored. Only recently, elegant neurophysiological studies have started to characterize the neural mechanisms of visual-vestibular integration during a heading discrimination task in awake macaque (Fetsch et al., 2013). These studies demonstrated that single neurons (Morgan et al., 2008) and populations of neurons (Fetsch et al., 2012) in the dorsal medial superior temporal area (dMST) integrated visual and vestibular information weighted by their reliability. However, the neural basis of reliability-weighted integration has yet to be identified in humans. Moreover, neurophysiological recordings focused selectively on dMST as one particular region of interest. Yet, over the past decade, evidence has accumulated showing that multisensory integration is not deferred until later processing stages in higher-order association cortices (Calvert et al., 2000; Beauchamp et al., 2004; Sadaghiani et al., 2009; Lewis and Noppeney, 2010; Werner and Noppeney, 2010), but starts already at the primary cortical level (Foxe et al., 2000; Ghazanfar and Schroeder, 2006; Bonath et al., 2007; Kayser et al., 2007; Lakatos et al., 2007; Lewis and Noppeney, 2010; Werner and Noppeney, 2010). These findings raise the question at which level of the cortical hierarchy sensory information is integrated weighted by their reliability in line with human behavior.

Traditionally, it is assumed that sensory signals that are close in time, space and structure are fused in a mandatory and automatic fashion into one unified percept (Jacobs, 1999; Ernst and Banks, 2002; van Beers et al., 2002; Battaglia et al., 2003; Knill and Saunders, 2003; Alais and Burr, 2004; Hillis et al., 2004; Saunders and Knill, 2004; Rosas et al., 2005; Fetsch et al., 2012). However, this classical forced fusion model has recently been challenged on two grounds. First, psychophysics studies have shown cases of only partial integration for signals that are spatiotemporally disparate (Wallace et al., 2004; Gepshtein et al., 2005; Parise et al., 2012). Moreover, recent EEG evidence (Donohue et al., 2011) indicated that multisensory integration depends on participant's attentional context indicating that multisensory integration is determined by both bottom-up sensory signals and top-down cognitive context (e.g., selective attention) (Alsius et al., 2005; Busse et al., 2005; Talsma et al., 2010).

This fMRI study combined psychophysics, multivariate decoding and quantitative predictions based on the MLE model (Ernst and Banks, 2002) to investigate the neural basis of bottom-up reliability-weighted integration and its interaction with top-down task-relevance in humans. In a spatial ventriloquist paradigm, participants were presented with auditory, visual and audiovisual spatially congruent and slightly disparate signals. Participants selectively reported the location of the visual or the auditory signal. From psychometric and neurometric functions we obtained quantitative predictions for the behavioral and neural weights based on the MLE model. Critically, imaging the entire auditory (Tian et al., 2001) and visual (Mishkin et al., 1983) spatial processing hierarchy enabled us to characterize the computational operations in low level auditory, visual and higher-order parietal cortices.

5.3 Materials and methods

Participants

After giving written informed consent, six healthy volunteers (2 female, mean age 28.8 years, range 22-36 years) participated in the fMRI study. All participants had normal or corrected-to normal vision and reported normal hearing. One participant was excluded due to excessive head motion (4.206 / 3.518 STD above the mean of the translational / rotational volume-wise head motion based on the included 5 participants). Data from the participants were also analyzed in chapter 3 and 4, except that in the current study we moreover analyzed data from unimodal conditions (see below). The study was approved by the human research review committee of the University of Tuebingen.

Stimuli

The visual stimulus was a cloud of 20 white dots (diameter: 0.43° visual angle) sampled from a bivariate Gaussian with a vertical standard deviation of 2.5° and a horizontal standard deviation of 2° or 14° (high and low visual reliability). The visual stimulus was presented on a black background (i.e., 100% contrast). The auditory stimulus was a burst of white noise with a 5ms on/off ramp. To create a virtual auditory spatial signal, the noise was convolved with spatially specific head-related transfer functions (HRTFs). The HRTFs were pseudo-individualized by matching participants' head width, heights, depth and circumference to the anthropometry of participants in the CIPIC database (Algazi et al., 2001) and were interpolated to the desired location of the auditory signal.

Experimental design and procedure

In unimodal conditions of the spatial ventriloquist paradigm, participants were presented either with auditory or with visual signals of low or high reliability. The signals were sampled from four possible locations along the azimuth (i.e., -10°, -3.3°, 3.3° or 10°). This yielded 4 (auditory locations) unimodal auditory and 4 (visual locations) x 2 (visual reliability: high vs. low) unimodal visual conditions. In bimodal conditions, participants were presented with synchronous auditory and visual signals of high and low visual reliability independently sampled from the four possible locations. This yielded 4 (auditory location) x 4 (visual location) x 2 (visual reliability) audiovisual conditions. For the current study, we only analyzed data from congruent ($\triangle$AV = 0°; 4 (signal location) x 2 (visual reliability) conditions) and slightly disparate conditions (3 (signal location) x 2 (visual reliability) conditions for $\triangle$AV = 6° and = -6°, respectively).

On each trial, spatial signals were presented for 50ms followed by a variable inter-stimulus fixation interval from 1.75-2.75s. Participants reported their auditory perceived location in the unisensory auditory and the audiovisual sessions with auditory report. They reported their visual perceived location in the visual and the audiovisual sessions with visual report. Participants indicated the perceived location by pushing one of four spatially corresponding buttons. Throughout the experiment, they fixated a central cross (1.6° diameter).

The subjects participated in 3-4 unimodal auditory, 3-4 unimodal visual and 20 bimodal sessions (10 auditory and 10 visual report; apart from one participant who performed 9 auditory and 11 visual report sessions). In each of the respective sessions we presented the 4 unimodal auditory conditions 88 times, the 8 unimodal visual conditions 44 times and the 32 audiovisual conditions 11 times. Further, 5.9% null-events were interspersed in the sequence of 352 stimuli per session. To maximize design efficiency, stimulus conditions were presented in a pseudorandomized fashion. We held the task (visual vs. auditory report) and bimodal versus unimodal conditions constant within a session and counterbalanced across sessions.

Experimental setup

Spatial stimuli were presented using Psychtoolbox 3.09 (www.psychtoolbox.org)(Brainard, 1997) running under MATLAB R2010a (MathWorks). Auditory stimuli were presented at ~75 dB SPL using MR-compatible headphones (MR Confon). Visual stimuli were back-projected onto a Plexiglas screen using an LCoS projector (JVC DLA-SX21). Participants viewed the screen through an extra-wide mirror mounted on the MR head coil resulting in a horizontal visual field of approx. 76° at a viewing distance of 26 cm. Participants performed the localization task using an MR-compatible custom-built button device. Participants' fixation was controlled by recording participants' pupil location using an MR-compatible custom-build infrared camera (sampling rate 50 Hz) mounted in front of the participants' right eye and iView software 2.2.4 (SensoMotoric Instruments). Analyses of this data showed that participants did not commit to condition-related eye movements (cf. last paragraph of results in chapter 4.4 and supplemental tab. S4.1).

Behavioral data

We discretized the four-level location reports to left versus right response. The fractions of right responses were plotted as a function of signal location for each stimulation x report condition (Fig. 5.1A-C). Because of the high number of trials in each participant, we individually fitted cumulative Gaussian functions to this data using maximum likelihood as implemented in Palamedes toolbox 1.5.0 (Prins and Kingdom, 2009). The Gaussians' mean (i.e., points of subject equality, PSE) and variance (σ) were used to compute the predicted visual weight (w_V in equation (1)), the predicted variance of the audiovisual percept (σ_{AV} in equation (2)), the empirical visual weight ($w_{V,emp}$ in equation (3); averaged for $\triangle AV = -6°$ and $+6°$) and the empirical unimodal and audiovisual variances (from congruent audiovisual conditions).

We employed 2 x 2 repeated measures ANOVAs to test the effects of visual reliability (high vs. low) and task (visual vs. auditory report) on the empirical visual weights and audiovisual variances (i.e., random effects analysis). We used paired t tests to compare the empirical visual weights and audiovisual variances against the MLE predictions and unimodal auditory and unimodal visual variances (Tab. 5.1). In the current study, results were deemed significant if $p < 0.05$.

MRI data acquisition

A 3T Siemens Magnetom Trio MR scanner was used to acquire both T1-weighted anatomical images and T2*-weighted axial echoplanar images (EPI) with BOLD contrast (gradient echo, parallel imaging using GRAPPA with an acceleration factor of 2, TR = 2480ms, TE = 40ms, flip angle=90°, FOV=192 mm×192 mm, image matrix 78×78, 42 transversal slices acquired interleaved in ascending direction, voxel size=2.5×2.5×2.5 mm + 0.25 mm interslice gap). In total, we acquired 353 volumes times 20 sessions for the bimodal conditions, 353 volumes times 6-8 sessions for the unimodal conditions, 161 volumes times 2-4 sessions for the auditory localizer and 159 volumes times 10-16 sessions for the visual retinotopic localizer (see below). This resulted in approximately 18 hours of scanning per participant assigned over 7-11 days. The first three volumes of each session were discarded to allow for T1 equilibration effects.

fMRI data analysis

Spatial ventriloquist paradigm

The fMRI data were analyzed with SPM8 (www.fil.ion.ucl.ac.uk/spm) (Friston et al., 1994). Scans from each participant were corrected for slice timing, were realigned and unwarped to correct for head motion and spatially smoothed with a Gaussian kernel of 3 mm FWHM. The time series in each voxel was high-pass filtered to 1/128 Hz. All data were analyzed in native subject space. The fMRI experiment was modeled in an event-related fashion with regressors entering into the design matrix after convolving each event-related unit impulse with a canonical hemodynamic response function and its first temporal derivative. In addition to modeling the 4 unimodal auditory, the 8 unimodal visual or the 32 audiovisual conditions in each session, the general linear models included the realignment parameters as nuisance covariates to account for residual motion artefacts. The factor task (visual vs. auditory report) was modeled across sessions. The parameter estimates pertaining to the canonical hemodynamic response function (HRF) defined the magnitude of the BOLD response to the unimodal or audiovisual stimuli in each voxel.

To apply MLE analysis to spatial representations at the neural level, we first extracted the parameter estimates of the HRF for each condition and session from voxels of regions defined in separate auditory and retinotopic localizer experiments (see below). This yielded activation patterns from the unimodal auditory and visual conditions and the bimodal congruent ($\triangle AV = 0°$) and slightly disparate ($\triangle AV \pm 6°$) audiovisual conditions. Individual activation patterns were z normalized to avoid the effects of image-wide activation changes. We then trained a linear support vector classification model (SVC, as implemented in LIBSVM 3.14(Chang and Lin, 2011)) to learn the mapping from activation patterns to the side (left vs. right) of the audiovisual signal. Importantly, we selectively used activation patterns from audiovisual *congruent* conditions from all but one audiovisual session for SVC training (i.e., training was done across sessions of auditory and visual report). The trained SVC model then decoded the signal side from the activation patterns of the spatially congruent and *disparate* audiovisual conditions of the remaining audiovisual session. In a leave-one-out cross-validation scheme, the training-test procedure was repeated for all audiovisual sessions. Finally, the SVC model was trained on audiovisual congruent conditions from all audiovisual sessions and then decoded the signal side from activation patterns of unimodal auditory and visual sessions.

Thus, the decoded signal sides represented auditory and visual spatial information at the neural level and were amenable to the same MLE analysis as the psychophysical localization responses (see above; Fig. 5.2). Due to the lower signal-to-noise ratio of fMRI compared to psychophysical data, we fitted neurometric functions to the proportion right decoded signals pooled across all participants (i.e., fixed effects analysis). Confidence intervals for empirical and predicted weights and variances were computed by using Palamedes' parametric bootstrap procedure (5000 bootstraps). We used two-tailed bootstrap tests (5000 bootstrap samples) (Efron and Tibshirani, 1994) to analyze the effects of visual reliability (high vs. low), task (visual vs. auditory report) and their interaction on the empirical visual weights and audiovisual variances and to compare those against the MLE predictions and unimodal auditory and visual variances (Tab. 5.1).

Auditory and visual retinotopic localizer

Regions of interest along the auditory and visual processing hierarchies were defined in a subject-specific fashion based on auditory and visual retinotopic localizers. In the auditory localizer, participants were presented with brief bursts of white noise at -10° or 10° angle (duration 500 ms, stimulus onset asynchrony 1 s). In a one-back task, participants indicated via a key press when the spatial location of the current trial was different from the previous trial. 20 s blocks of auditory stimulation (i.e., 20 trials) alternated with 13 s of fixation periods. The auditory locations were presented in a pseudorandomized fashion to optimize design efficiency. Similar to the main experiment, the auditory localizer sessions were modeled in an event-related fashion. Auditory responsive regions were defined as voxels in superior temporal and Heschl's gyrus showing significant activations for auditory stimulation relative to fixation ($p < 0.05$, family-wise error corrected). Within these regions, we defined primary auditory cortex (A1) based on cytoarchitectonic probability maps (Eickhoff et al., 2005) and referred to the remainder (i.e., planum temporale and posterior superior temporal gyrus) as higher order auditory cortex (hA).

Visual regions of interest were defined using standard phase-encoded retinotopic mapping (Sereno et al., 1995). Participants viewed a checkerboard background flickering at 7.5 Hz through a rotating wedge aperture of 70° width (polar angle mapping) or an expanding/contracting ring (eccentricity mapping). The periodicity of the apertures was 42s. Visual responses were modeled by entering a sine and cosine convolved with the hemodynamic response function as regressors into the design matrix of the general linear model. The preferred polar angle (or eccentricity, respectively) was determined as the phase lag for each voxel by computing the angle between the parameter estimates for the sine and the cosine. The phase lags for each voxel were projected on the reconstructed, inflated cortical surface using Freesurfer 5.1.0 (Dale et al., 1999). Visual regions V1-V3 and IPS0-IPS4 were defined as phase reversal in angular retinotopic maps. IPS0-4 were defined as phase reversal along the anatomical IPS resulting in contiguous, approximately rectangular regions (Swisher et al., 2007).

For the decoding analyses, the auditory and visual regions were combined from the left and right hemisphere. SVC training was then applied separately to activation patters from each region. To improve the signal-to-noise ratio when fitting neurometric functions (cf. Fig. 5.2), the decoded signal sides from low-level visual regions (V1-3), intraparietal sulcus (IPS0-4) and low-level auditory regions (A1, hA) regions were pooled. Additional analyses showed similar audiovisual spatial integration within these three regions.

5.4 Results

Spatial ventriloquist paradigm

In the fMRI study, five participants were presented with auditory, visual and audiovisual signals sampled randomly from four possible spatial locations along the azimuth (i.e., -10°, -3.3°, 3.3° or 10°). Audiovisual signals were either spatially congruent (△AV = 0°) or slightly disparate (△AV = ±6°). The reliability of the visual signal could be high or low. Participants reported their auditory perceived location in the unisensory auditory and the audiovisual sessions with auditory report. They report their visual perceived location in the visual and the audiovisual sessions with visual report.

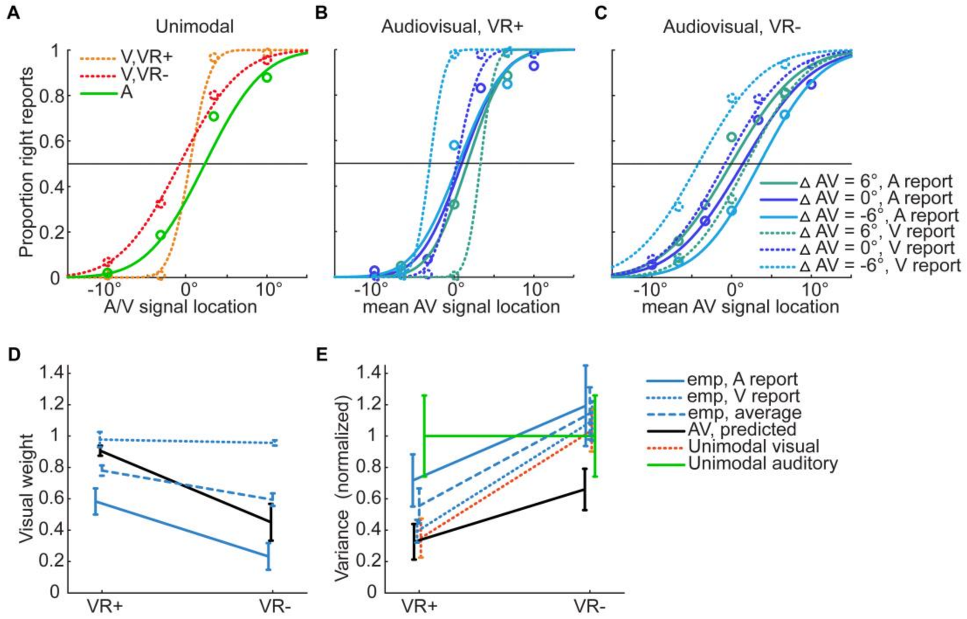


Figure 5.1. MLE analysis of psychophysical data. (A) Psychometric functions were fitted to the proportion right reports plotted as a function of the signal location from unimodal auditory (A) and visual conditions of high (V, VR+) and low (V, VR-) visual reliability. **(B, C)** In bimodal conditions, psychometric functions were fitted to proportion right reports plotted as a function of the mean audiovisual (AV) signal location. Data was fitted separately for congruent (△AV = 0°; △AV = A - V) and disparate conditions (△AV = ±6°), conditions of high (B) versus low (C) visual reliability and auditory versus visual report. **(D)** Predicted (equation (1)) and empirical (equation (3)) visual weights (mean ± SEM across participants) for high versus low visual reliability and visual versus auditory report. For illustration, the average of the empirical weights across the latter two conditions is plotted. **(E)** Unimodal and audiovisual predicted (equation (2)) and empirical variances (mean ± SEM across participants) for the same combination of conditions as in (D). Variances were normalized by the auditory variance.

Key predictions of the MLE model

Under the classical forced fusion assumption, the MLE model makes two key quantitative predictions for participants' spatial estimates in audiovisual conditions: First, the most reliable unbiased estimate of the object's location ($\hat{S}_{AV}$) is obtained by combining the auditory (X_A) and visual (X_V) perceived locations in proportion to their relative reliability (r_A, r_V, i.e., the inverse of the variance, $r = 1/\sigma^2$) as obtained from the unisensory conditions:

$$(1) \qquad \hat{S}_{AV} = w_A X_A + w_V X_V \quad \text{with} = \; w_A = \frac{r_A}{r_A + r_V} \text{ and } w_V \; = \frac{r_V}{r_A + r_V}$$

Second, multisensory integration reduces the variance of the audiovisual estimate (σ_{AV}^2) as compared to the unimodal variances (σ_A^2, σ_V^2):

$$(2) \qquad \sigma_{AV}^2 = \frac{\sigma_A^2 \sigma_V^2}{\sigma_A^2 + \sigma_V^2}$$

MLE analysis of psychophysics data

Figure 5.1A-C shows the proportion 'right' responses as a function of true signal location and the corresponding fitted cumulative Gaussians for each stimulation x report condition (see above). For each cumulative Gaussian we obtained the variance and its mean (i.e., the point of subjective equality (PSE) that is the abscissa for 50% proportion 'right' responses).

First, we investigated whether participants integrated audiovisual signals weighted by their reliability as predicted by MLE. The variances of the cumulative Gaussians for the unisensory visual and auditory conditions were used to determine the 'optimal' weights that participants should apply to the bimodal visual and auditory signals (equation 1). The empirical weights were computed from the PSE of the psychometric functions of the audiovisual conditions according to the following equation (Fetsch et al., 2012):

$$(3) \qquad w_{V,\,emp} = \frac{\mathrm{PSE}_{\Delta\,AV\,=\,\pm 6°} - \mathrm{PSE}_{\Delta\,AV\,=\,0°} + \frac{\Delta\,AV}{2}}{\Delta\,AV}$$

As shown by lateral shifts of the PSEs in Fig. 5.1B and C, during the cue conflict conditions the perceived auditory location shifted from the congruent audiovisual location towards the true visual location when visual reliability was high, but towards the true auditory location when visual reliability was low. By contrast, the perceived visual location was shifted towards the visual location for both high and low visual reliabilities. This pattern in perceptual bias was also reflected in the sensory weights (Fig. 5.1D): During auditory report conditions, the visual weight was greater than 0.5 for high visual reliability ($w_{V,\,emp} = 0.583$), but smaller than 0.5 for low visual reliability ($w_{V,\,emp} = 0.233$). By contrast, during visual report conditions, the visual weight was always close to 1 ($w_{V,\,emp} = 0.937$) indicating that the auditory influence on the perceived visual location was statistically significant ($w_{A, emp} = 1 - w_{V, emp} = 0.063$, $p = 0.001$, one-sample t test against 0), but very small. Thus, the visual weights violated the MLE predictions (Tab. 5.1). Only when averaged across auditory and visual report, the visual weights followed the MLE predictions in a qualitative fashion (Fig. 5.1D). However, a 2 (reliability: high vs. low) x 2 (task: auditory vs. visual report) repeated measures ANOVA identified a main effect of reliability ($F_{1,4} = 28.237$, $p = 0.006$), task context ($F_{1,4} = 36.496$, $p = 0.004$), and their interaction ($F_{1,4} = 9.174$, $p = 0.039$) indicating that reliability-driven reweighting determined integration only in case of auditory report.

Table 5.1. Statistical comparison of empirical variances ($\sigma_{AV,emp}^2$) and weights ($w_{V,emp}$) from the four bimodal conditions against the predictions ($\sigma_{AV,pred}^2$, $w_{V,pred}$) and unimodal variances (σ_{uniV}^2, σ_{uniA}^2).

	VR+, A report	VR-, A report	VR+, V report	VR-, V report
	$\sigma_{AV,emp}^2$ - $\sigma_{AV,pred}^2$			
Psychophysics	0.023	0.031	0.361	0.036
IPS0-4	0.015	0.408	0.304	0.154
	$\sigma_{AV,emp}^2$ - σ_{uniV}^2			
Psychophysics	0.031	0.530	0.599	0.526
IPS0-4	0.006	0.973	0.120	0.271
	$\sigma_{AV,emp}^2$ - σ_{uniA}^2			
Psychophysics	0.098	0.249	0.037	0.725
IPS0-4	0.004	0.051	0.005	0.394
	$w_{V,emp} - w_{V,pred}$			
Psychophysics	0.029	0.063	0.036	0.013
IPS0-4	0.226	0.066	0.051	0.463

Note: Numbers denote p values. Psychophysical parameters were compared using two-tailed paired t tests on individual parameters (random-effects analysis, df = 4). Parameters from IPS0-4 were compared using a two-tailed bootstrap test (5000 bootstraps) on parameters computed across the sample (fixed-effects analysis). A = auditory, V = visual, VR+/- = High / low visual reliability.

Second, we investigated whether multisensory integration reduced the variance of the spatial estimate as predicted by MLE (equation 2). To maximize the effect of multisensory integration, we limited this analysis to the congruent trials only. As shown in Fig. 5.1E, even for congruent trials the variances of the audiovisual percepts were significantly greater than predicted by MLE in most stimulation and report conditions (Tab.

5.1). Specifically, the variance of the perceived visual location during audiovisual stimulation was comparable to the unisensory visual variance ($p > 0.05$). Likewise, the variance of the perceived auditory location during audiovisual stimulation was not smaller than unisensory visual variance, though it was smaller than the unisensory auditory variance if visual reliability was high ($p = 0.049$, one-tailed paired t test). Moreover, in a 2 (reliability: high vs. low) x 2 (task: auditory vs. visual report) repeated measures ANOVA we observed a main effect of reliability ($F_{1,4} = 28.5468$, $p = 0.006$; effect of task and the interaction were not significant, $p > 0.05$).

Collectively, our behavioral results suggested that auditory and visual signals were only partially integrated proportional to reliability dependent on the modality-specific report. Even spatiotemporally congruent auditory and visual signals were not fully fused into one unified percept as predicted by the MLE model.

MLE analysis of fMRI data

To investigate the neural processes underlying multisensory spatial integration at the psychophysical level, we decoded spatial information from fMRI activation patterns. The patterns were recorded from low-level visual regions (V1-V3), intraparietal sulcus (IPS0-4) and low-level auditory regions (primary auditory cortex and planum temporale, lA). Using the fMRI activation patters selectively from audiovisual congruent conditions ($\triangle AV = 0°$), we trained a support-vector classification model to learn the mapping from activation patterns to the side of the signal (left vs. right). The trained model then decoded the signal side from activation pattern in disparate audiovisual conditions (i.e., $\triangle AV = \pm 6°$) and unimodal auditory and visual conditions. Thus, the decoded signal sides represented auditory and visual spatial information at the neural level and were amenable to the same MLE analysis as the psychophysical location reports (cf. Fig. 5.2).

Among the regions in the auditory (Tian et al., 2001) and visual (Mishkin et al., 1983) spatial processing hierarchy, selectively IPS0-4 demonstrated significant reliability-driven reweighting of audiovisual signals as observed at the psychophysical level (Fig. 5.2D; $w_{V,\,emp} = 0.915$ for high and $w_{V,\,emp} = 0.655$ for low visual reliability; effect of visual reliability, $p = 0.014$, bootstrap test). Further, visual report ($w_{V,\,emp} = 1.019$) increased the visual weight relative to an auditory report ($w_{V,\,emp} = 0.674$; effect of task, $p = 0.001$; the interaction was not significant, $p > 0.05$). The empirical and the predicted weights were statistically indistinguishable (cf. Tab. 5.1). In parallel to the psychophysical results, however, the empirical weights were much closer to the predicted weights if we averaged over conditions of auditory and visual report (see Fig. 5.2D).

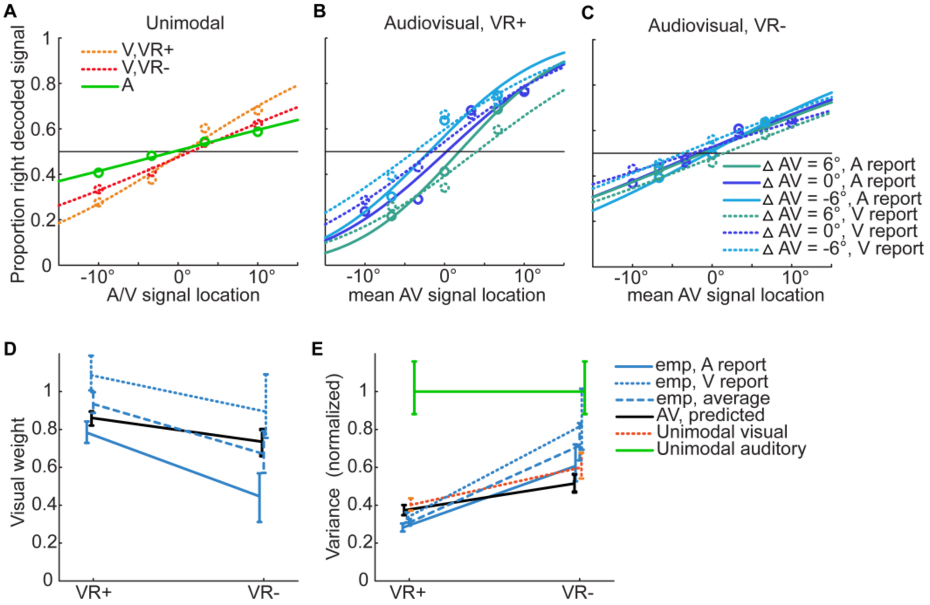


Figure 5.2. MLE analysis of fMRI data in intraparietal sulcus (IPS0-4). (A) In IPS0-4, neurometric functions were fitted to the proportion right decoded signals plotted as a function of signal location from unimodal auditory (A) and visual conditions of high (V, VR+) and low (V, VR-) visual reliability. **(B, C)** In bimodal conditions, neurometric functions were fitted to proportion right decoded signals plotted as function of the mean audiovisual (AV) signal location. Data was fitted separately for congruent (∆AV = 0°) and disparate conditions (∆AV = ±6°; ∆AV = A - V), high (B) versus low (C) visual reliability and auditory versus visual selective report. **(D)** Predicted (equation (1)) and empirical (equation (3)) visual weights (mean and 68% bootstrapped confidence interval) for high versus low visual reliability and visual versus auditory-selective report. For illustration, the average of the empirical weights across the latter two conditions is plotted. **(E)** Unimodal and audiovisual predicted (equation (2)) and empirical variances (mean and 68% bootstrapped confidence interval) for the same combination of conditions as in (D). Variances were normalized by the auditory variance.

The variance of spatial representation of unimodal visual signals in IPS0-4 was much lower than for unimodal auditory signals even if visual reliability was low (Fig. 5.2E). Therefore, the MLE model did not predict a strong reduction of audiovisual variance beyond the unimodal visual variance (i.e., the predicted reduction is maximal in case of equal reliability). Accordingly, the empirical matched the predicted audiovisual variance (Tab. 5.1), even though the empirical audiovisual variance was not significantly smaller than the unimodal visual variance. However, highly reliable visual signals significantly reduced the audiovisual variance compared to unimodal auditory variance. Interestingly, the audiovisual variance was significantly smaller than the unimodal visual, auditory and even the predicted audiovisual variance if visual reliability was high and participants reported the auditory signals. As expected, visual signals of low reliability led to higher variance than highly reliable signals (effect of visual reliability, p = 0.027, bootstrap test; effect of task and the interaction were not significant, p > 0.05).

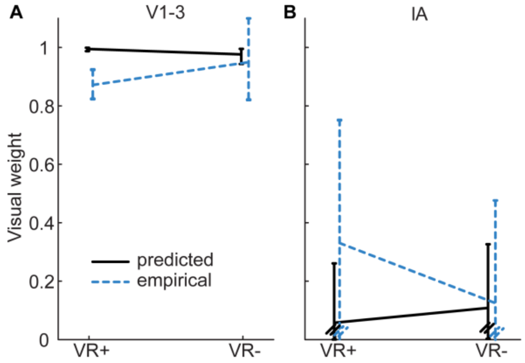


Figure 5.3. Predicted and empirical visual weights resulting from the MLE analysis of fMRI data in low-level visual (V1-3) and low-level auditory (lA) regions. (A) Visual Weights derived from fMRI data in V1-3. **(B)** Visual weights derived from fMRI data in lA. Note that the weights are pooled over visual versus auditory-selective report.

In contrast to IPS0-4, low-level visual and auditory regions gave a large weight to the spatial information of their preferred modality (V1-3, $w_{V, emp}$ = 0.882; lA, $w_{V, emp}$ = 0.232; Fig. 5.3). Critically, the spatial representations in these regions were unaffected by visual reliability, the task and the interaction of both factors ($p > 0.05$, bootstrap test). However, consistent with reports of multisensory integration at low levels of the processing hierarchy (Foxe et al., 2000; Bonath et al., 2007; Kayser et al., 2007; Lakatos et al., 2007; Lewis and Noppeney, 2010; Werner and Noppeney, 2010), visual regions integrated auditory ($p < 0.001$, two-tailed bootstrap test on $w_{V, emp}$ against 1) and auditory regions integrated visual spatial information ($p = 0.032$, one-tailed bootstrap test on $w_{V, emp}$ against 0).

In summary, analyses of fMRI data revealed that specifically IPS weighted audiovisual signals by their reliability and prioritized signals of the reported modality. In parallel to the behavioral results, the MLE predictions were consistently violated because IPS did not fully fuse audiovisual signals into a unified spatial representation.

5.5 Discussion

Psychophysics studies have demonstrated that human observers combine signals within and across the senses weighted in proportion to their reliability as predicted by the MLE model (Jacobs, 1999; Ernst and Banks, 2002; van Beers et al., 2002; Battaglia et al., 2003; Knill and Saunders, 2003; Alais and Burr, 2004; Hillis et al., 2004; Saunders and Knill, 2004; Rosas et al., 2005). Combining classical MLE analysis with a multivariate fMRI-decoding approach, we show that selectively IPS computes reliability-weighted audiovisual spatial estimates, in parallel to psychophysical results. However, the visual weights violated the MLE predictions and were larger for visual than auditory report. Earlier regions of the auditory (Tian et al., 2001) and visual (Mishkin et al., 1983) spatial processing hierarchy represented spatial signals of their preferred and, slightly, non-preferred modality, but the representations did not depend on sensory reliability or the modality of report.

Reliability-weighted integration

Referring to a maximum-likelihood criterion, reliability-weighted integration of noisy signals is the optimal strategy to estimate an uncertain physical quantity. Numerous psychophysical studies showed that humans integrate unisensory (Jacobs, 1999; Knill and Saunders, 2003; Hillis et al., 2004), multisensory (Ernst and Banks, 2002; Battaglia et al., 2003; Alais and Burr, 2004; Rosas et al., 2005) and motor-related signals (van Beers et al., 2002; Saunders and Knill, 2004) in this fashion. Yet, evidence that the brain applies reliability-driven integration has been rare except for recent evidence of visuo-vestibular integration in monkeys' dMST region (Morgan et al., 2008; Fetsch et al., 2012; Fetsch et al., 2013). This constitutes a serious empirical gap in Bayesian theories of mind and brain which rest on the assumption that the brain represents uncertainty (Knill and Pouget, 2004). Here, by using neurometric functions and multivariate decoding, we demonstrate that human IPS weighs audiovisual signals proportional to relative sensory reliability (Fig. 5.2D). Because we unpredictably manipulated visual reliability in each trial, the current finding shows that the brain represents a signals' uncertainty (i.e., the inverse of reliability) automatically in parallel to the signals' value per se. Thus, the brain estimates physical quantities in a probabilistic fashion (Knill and Pouget, 2004). Because our multivariate decoding approach rests on large-scale population responses, the current results are consistent with the notion that probabilistic population codes implement such probabilistic computations (Ma et al., 2006). However, reliability-weighted integration in IPS was suboptimal as compared to the MLE predictions.

Task-dependent deviations from optimal weighting

Audiovisual integration at the psychophysical level and in IPS depended on visual reliability as well as the modality of report (Fig. 5.1D, 5.2D): The signals of the task-relevant modality obtained larger weights. Thus, the weighting of the signals only approximated the optimal weighting suggested by the MLE model if we pooled across the conditions of auditory and visual report. The influence of the modality of report revealed that the participants did not integrate the signals into a unified, task-independent spatial representation as predicted by the MLE model.

In contrast to our study, classical studies on the MLE model (Jacobs, 1999; Ernst and Banks, 2002; van Beers et al., 2002; Battaglia et al., 2003; Knill and Saunders, 2003; Alais and Burr, 2004; Hillis et al., 2004; Saunders and Knill, 2004; Rosas et al., 2005; Fetsch et al., 2012) encourage observers to focus equally on signals of both modalities to emphasize a 'forced' fusion of the signals. The current results suggest that only such an integrative focus (or, in approximation, an 'averaged' focus) leads to MLE-consistent weighting. However, if observers do not commit to the integrative focus, for example if the spatiotemporal conflict between the signals becomes too large (Gepshtein et al., 2005; Parise et al., 2012), the predictions of the MLE model are violated.

In line with this observation, a recent Bayesian model of causal inference (Kording et al., 2007; Shams and Beierholm, 2010) suggests that the signals' relative weights shift in direction of the task-relevant signal if large signal discrepancies lead to the inference that the signals were caused by independent sources. Accordingly, a larger weight for the task-relevant signal in IPS demonstrated that the brain infers, despite the signals' small spatial disparity in the current study (i.e., ± 6°), that the signals were not caused by the same source. Hence, IPS performs computations consistent with causal inference.

Lack of integration benefits

Consistent with the finding that participants did not integrate the signals into a unified, task-independent spatial representation, we did not find evidence that the variance of the spatial estimates benefits from signal integration as predicted by the MLE model. This finding arises from the non-optimal weighting of the signals, but it might also arise due to methodological reasons: In IPS, the unimodal visual variance was generally much lower than the auditory variance even in case of low visual reliability (Fig. 5.2E). Thus, the predicted integration benefit might be too small to be detectable (note that equation (2) predicts a maximum benefit if the unimodal variances are equal). However, at the psychophysical level a considerable integration benefit was predicted for low visual reliability (Fig. 5.1E). Because the participants gave the visual signals too much weight in case of visual and too little weight in case of auditory report compared to the MLE predictions (Fig. 5.1D), they did not benefit from the spatial information available in both modalities as predicted. In other terms, the participants segregated spatial information from the task-irrelevant modality, presumably because they did not infer a common source of the signals as presupposed by the MLE model.

Multisensory integration along the spatial processing hierarchy

By imaging the entire spatial processing hierarchy, we found distinct multisensory processes at different hierarchical levels: Consistent with previous evidence showing that multisensory integration starts already at the primary cortical level (Foxe et al., 2000; Ghazanfar and Schroeder, 2006; Bonath et al., 2007; Kayser et al., 2007; Lakatos et al., 2007; Lewis and Noppeney, 2010; Werner and Noppeney, 2010), we found auditory influences on low-level visual and visual influences on low-level auditory regions. Critically, the multisensory processes of reliability-weighting and causal inference were restricted to higher multisensory association cortex. Thus, our results show that it is more crucial to characterize specific multisensory processes at different stages of the cortical hierarchies than to investigate unspecific multisensory integration per se. Eventually, many regions of neocortex might have access to some form of multisensory information (Ghazanfar and Schroeder, 2006) via top-down influences from higher-order association regions (Macaluso et al., 2000; Macaluso and Driver, 2005), direct cortico-cortical connectivity (Falchier et al., 2002) or feed-forward thalamic mechanisms (Lakatos et al., 2007).

In conclusion, the current study demonstrates for the first time that human multisensory association cortex implements probabilistic computations to model an uncertain multisensory environment. However, the probabilistic computations are more complex than ‘mandatory’ reliability-weighted integration of signals because they account for causal inferences on the signals’ origin.

5.6 Acknowledgments

This study was funded by the Max Planck Society and the European Research Council. We thank P. Ehses for advice on the fMRI sequence and S. Schwarzkopf for providing stimulus scripts for retinotopic mapping (http://sampendu.wordpress.com/retinotopy-tutorial/).

5.7 References

Alais D, Burr D (2004) The ventriloquist effect results from near-optimal bimodal integration. Curr Biol 14:257-262.

Algazi VR, Duda RO, Thompson DM, Avendano C (2001) The cipic hrtf database. In: Applications of Signal Processing to Audio and Acoustics, 2001 IEEE Workshop on the, pp 99-102: IEEE.

Alsius A, Navarra J, Campbell R, Soto-Faraco S (2005) Audiovisual integration of speech falters under high attention demands. Curr Biol 15:839-843.

Battaglia PW, Jacobs RA, Aslin RN (2003) Bayesian integration of visual and auditory signals for spatial localization. J Opt Soc Am A Opt Image Sci Vis 20:1391-1397.

Beauchamp MS, Argall BD, Bodurka J, Duyn JH, Martin A (2004) Unraveling multisensory integration: patchy organization within human STS multisensory cortex. Nat Neurosci 7:1190-1192.

Bonath B, Noesselt T, Martinez A, Mishra J, Schwiecker K, Heinze HJ, Hillyard SA (2007) Neural basis of the ventriloquist illusion. Curr Biol 17:1697-1703.

Brainard DH (1997) The psychophysics toolbox. Spatial vision 10:433-436.

Busse L, Roberts KC, Crist RE, Weissman DH, Woldorff MG (2005) The spread of attention across modalities and space in a multisensory object. Proc Natl Acad Sci U S A 102:18751-18756.

Calvert GA, Campbell R, Brammer MJ (2000) Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. Curr Biol 10:649-657.

Chang CC, Lin CJ (2011) LIBSVM: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology (TIST) 2:27.

Dale AM, Fischl B, Sereno MI (1999) Cortical surface-based analysis. I. Segmentation and surface reconstruction. Neuroimage 9:179-194.

Donohue SE, Roberts KC, Grent-'t-Jong T, Woldorff MG (2011) The cross-modal spread of attention reveals differential constraints for the temporal and spatial linking of visual and auditory stimulus events. J Neurosci 31:7982-7990.

Efron B, Tibshirani RJ (1994) An introduction to the bootstrap. London: Chapmann and Hall.

Eickhoff SB, Stephan KE, Mohlberg H, Grefkes C, Fink GR, Amunts K, Zilles K (2005) A new SPM toolbox for combining probabilistic cytoarchitectonic maps and functional imaging data. Neuroimage 25:1325-1335.

Ernst MO, Banks MS (2002) Humans integrate visual and haptic information in a statistically optimal fashion. Nature 415:429-433.

Falchier A, Clavagnier S, Barone P, Kennedy H (2002) Anatomical evidence of multimodal integration in primate striate cortex. J Neurosci 22:5749-5759.

Fetsch CR, Deangelis GC, Angelaki DE (2013) Bridging the gap between theories of sensory cue integration and the physiology of multisensory neurons. Nat Rev Neurosci 14:429-442.

Fetsch CR, Pouget A, DeAngelis GC, Angelaki DE (2012) Neural correlates of reliability-based cue weighting during multisensory integration. Nat Neurosci 15:146-154.

Foxe JJ, Morocz IA, Murray MM, Higgins BA, Javitt DC, Schroeder CE (2000) Multisensory auditory-somatosensory interactions in early cortical processing revealed by high-density electrical mapping. Brain Res Cogn Brain Res 10:77-83.

Friston KJ, Holmes AP, Worsley KJ, Poline JP, Frith CD, Frackowiak RSJ (1994) Statistical parametric maps in functional imaging: a general linear approach. Human brain mapping 2:189-210.

Gepshtein S, Burge J, Ernst MO, Banks MS (2005) The combination of vision and touch depends on spatial proximity. J Vis 5:1013-1023.

Ghazanfar AA, Schroeder CE (2006) Is neocortex essentially multisensory? Trends Cogn Sci 10:278-285.

Hillis JM, Watt SJ, Landy MS, Banks MS (2004) Slant from texture and disparity cues: optimal cue combination. J Vis 4:967-992.

Jacobs RA (1999) Optimal integration of texture and motion cues to depth. Vision Res 39:3621-3629.

Kayser C, Petkov CI, Augath M, Logothetis NK (2007) Functional imaging reveals visual modulation of specific fields in auditory cortex. J Neurosci 27:1824-1835.

Knill DC, Saunders JA (2003) Do humans optimally integrate stereo and texture information for judgments of surface slant? Vision Res 43:2539-2558.

Knill DC, Pouget A (2004) The Bayesian brain: the role of uncertainty in neural coding and computation. Trends Neurosci 27:712-719.

Kording KP, Beierholm U, Ma WJ, Quartz S, Tenenbaum JB, Shams L (2007) Causal inference in multisensory perception. PLoS One 2:e943.

Lakatos P, Chen CM, O'Connell MN, Mills A, Schroeder CE (2007) Neuronal oscillations and multisensory interaction in primary auditory cortex. Neuron 53:279-292.

Lewis R, Noppeney U (2010) Audiovisual synchrony improves motion discrimination via enhanced connectivity between early visual and auditory areas. J Neurosci 30:12329-12339.

Ma WJ, Beck JM, Latham PE, Pouget A (2006) Bayesian inference with probabilistic population codes. Nat Neurosci 9:1432-1438.

Macaluso E, Driver J (2005) Multisensory spatial interactions: a window onto functional integration in the human brain. Trends Neurosci 28:264-271.

Macaluso E, Frith CD, Driver J (2000) Modulation of human visual cortex by crossmodal spatial attention. Science 289:1206-1208.

Mishkin M, Ungerleider LG, Macko KA (1983) Object vision and spatial vision: two cortical pathways. Trends in neurosciences 6:414-417.

Morgan ML, Deangelis GC, Angelaki DE (2008) Multisensory integration in macaque visual cortex depends on cue reliability. Neuron 59:662-673.

Parise CV, Spence C, Ernst MO (2012) When correlation implies causation in multisensory integration. Curr Biol 22:46-49.

Prins N, Kingdom FAA (2009) Palamedes: Matlab Routines for Analyzing Psychophysical In.

Rosas P, Wagemans J, Ernst MO, Wichmann FA (2005) Texture and haptic cues in slant discrimination: reliability-based cue weighting without statistically optimal cue combination. J Opt Soc Am A Opt Image Sci Vis 22:801-809.

Sadaghiani S, Maier JX, Noppeney U (2009) Natural, metaphoric, and linguistic auditory direction signals have distinct influences on visual motion processing. J Neurosci 29:6490-6499.

Saunders JA, Knill DC (2004) Visual feedback control of hand movements. J Neurosci 24:3223-3234.

Sereno MI, Dale AM, Reppas JB, Kwong KK, Belliveau JW, Brady TJ, Rosen BR, Tootell RB (1995) Borders of multiple visual areas in humans revealed by functional magnetic resonance imaging. Science 268:889-893.

Shams L, Beierholm UR (2010) Causal inference in perception. Trends Cogn Sci 14:425-432.

Swisher JD, Halko MA, Merabet LB, McMains SA, Somers DC (2007) Visual topography of human intraparietal sulcus. J Neurosci 27:5326-5337.

Talsma D, Senkowski D, Soto-Faraco S, Woldorff MG (2010) The multifaceted interplay between attention and multisensory integration. Trends Cogn Sci 14:400-410.

Tian B, Reser D, Durham A, Kustov A, Rauschecker JP (2001) Functional specialization in rhesus monkey auditory cortex. Science 292:290-293.

van Beers RJ, Wolpert DM, Haggard P (2002) When feeling is more important than seeing in sensorimotor adaptation. Curr Biol 12:834-837.

Wallace MT, Roberson GE, Hairston WD, Stein BE, Vaughan JW, Schirillo JA (2004) Unifying multisensory signals across time and space. Exp Brain Res 158:252-258.

Werner S, Noppeney U (2010) Distinct functional contributions of primary sensory and association areas to audiovisual integration in object categorization. J Neurosci 30:2662-2675.

6 Bayesian learning of sensory reliability in multisensory perception

6.1 Abstract

It is unknown how the brain represents sensory reliability which is crucial to perceive physical quantities in a Bayesian fashion. Here, we show that human observers weight audiovisual signals proportional to posterior visual reliability which they learned, consistent with Bayesian inference, by combining reliability from prior and present signals. The result suggests that the brain uses Bayesian inference to link perception and sensory learning.

6.2 Introduction

When we perceive our environment, such as locating a bouncing tennis ball, our brain has to infer physical properties from noisy, unreliable sensory signals (Faisal et al., 2008). To optimally estimate a physical quantity given the signals' dynamically changing reliability, Bayesian theory suggests combining prior knowledge with new evidence provided by the signals (Yuille and Buelthoff, 1996; Knill and Pouget, 2004). For such a Bayesian inference, the prior‘s as well as the signals' reliability have a key role: The prior and the signals are weighted proportional to their reliability to optimally reduce the error in the estimate. For example, when not clearly seeing or hearing a tennis ball, we could only a priori assume that it likely bounces within the playing field, but when seeing and hearing the ball, we would locate it giving the more reliable of both signals a stronger weight. In face of very unreliable signals, human observers indeed rely on their priors (Kording et al., 2004; Kording and Wolpert, 2004; Berniker et al., 2010), and if multiple signals are relatively certain, they rely on the more certain signal (Ernst and Banks, 2002; Battaglia et al., 2003; Knill and Saunders, 2003; Alais and Burr, 2004). Thus, Bayesian inference in perception crucially requires the brain to represent the prior's as well as the signals' reliability.

However, while human observers learn the prior's reliability by extensive training (Kording et al., 2004; Kording and Wolpert, 2004; Berniker et al., 2010), it is unknown whether the brain represents sensory reliability immediately from single signals or whether it learns sensory reliability also from past signals. A probabilistic population code could represent a signal's reliability instantaneously via the gain of the neurons' population response to the signal (Ma et al., 2006), without the need for learning reliability from past signals. However, in natural environments signal reliability changes systematically, for example when a tennis ball's seen location becomes increasingly unreliable during sunset. In this case, a Bayesian learner would optimally learn this statistical regularity by updating prior knowledge on sensory reliability obtained from past signals with new evidence from current signals.

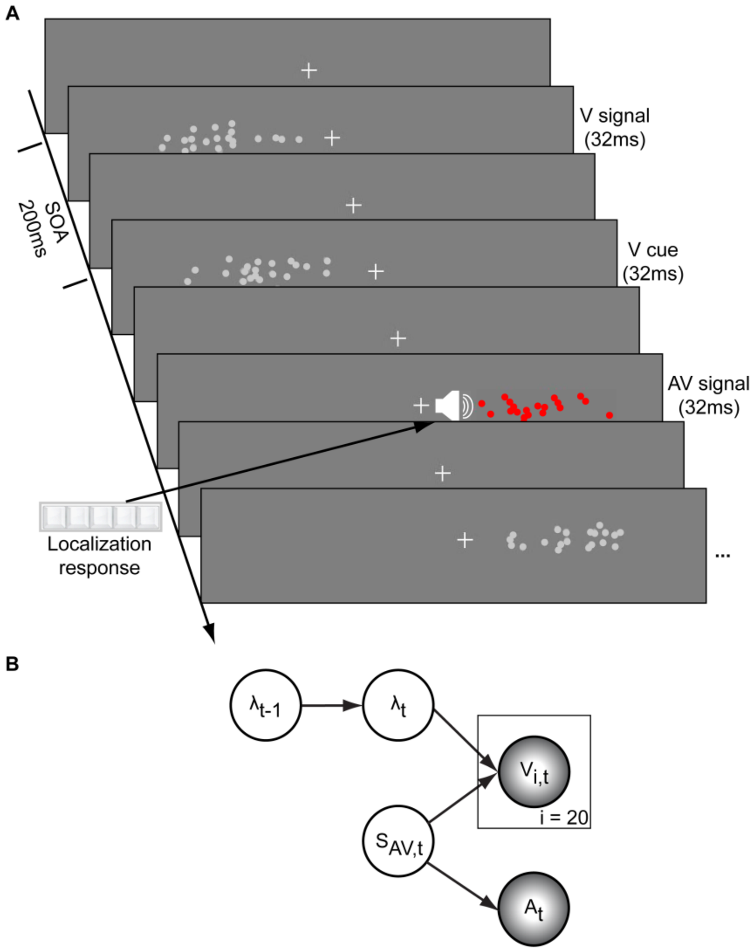


Figure 6.1. Spatial ventriloquist paradigm and generative Bayesian model for learning visual reliability. (**A**) Visual (V) signals (20 bright dots) were presented at 5 Hz (i.e., a stimulus onset asynchrony (SOA) of 200 ms). The clouds' variance (i.e., inverse of visual reliability) was temporally manipulated between 2° and 18° STD according to a sinusoid or two random walks (cf. fig. 6.2). The cloud's location mean was independently resampled from five possible locations (-10°, -5°, 0°, 5°, 10°) at a SOA jittered between 1.4 and 2.8 s. In synchrony with the change in the cloud's location, the dots changed their colour and a sound was presented (AV signal). Participants localized the sound using five response buttons. The location of the sound was sampled from the two possible locations adjacent to the visual cloud's mean location (i.e., ± 5° AV spatial discrepancy). (**B**) The generative Bayesian model for the Bayesian learner assumes that an audiovisual source ($S_{AV,t}$) creates visual ($V_{i,t}$) and auditory (A_t) spatial signals. Importantly, the reliability (i.e., 1/variance) of the visual signal at time t (λ_t) is estimated by updating prior information on reliability from previous visual signals (λ_{t-1}).

6.3 Materials and methods

Participants

56 healthy volunteers participated in the study after giving written informed consent (28 female, mean age 26.6 years, range 18-52 years). All participants were naïve to the purpose of the study. All participants had normal or corrected-to-normal vision and reported normal hearing. The study was approved by the human research review committee of the University of Tuebingen.

Stimuli

The visual spatial stimulus was a Gaussian cloud of twenty bright grey dots (0.56° diameter, vertical standard deviation 1.5°, luminance 106 cd/m^2) presented on a dark grey background (luminance 62 cd/m^2, i.e., 71% contrast). The location and the reliability of the visual signal were manipulated by the mean and the horizontal standard deviation of the Gaussian cloud of dots (see below). The auditory spatial cue was a burst of white noise with a 5 ms on/off ramp. To create a virtual auditory spatial cue, the noise was convolved with spatially specific head-related transfer functions (HRTFs). The HRTFs were pseudo-individualized by matching participants' head width, heights, depth and circumference to the anthropometry of subjects in the CIPIC database (Algazi et al., 2001). HRTFs from the available locations in the database were interpolated to the desired locations of the auditory cue.

Experimental design and procedure

In a spatial ventriloquist paradigm, participants were presented with a sequence of Gaussian clouds of dots at a rate of 5 Hz (Fig. 6.1A). The cloud's standard deviation changed according to a i. sinusoidal sequence, ii. random walk sequence 1 or iii. random walk sequence 2:

i. *Sinusoidal sequence (Sinus):* A sinusoidal sequence was generated with a period of 30s (initially at the monitor's refresh rate of 60Hz but then subsampled to 5Hz). Across participants, the starting phase of the sequence was randomized. During the ~65 min of the experiment, each participant completed ~ 130 cycles of the sinusoidal sequence.

ii. *Random walk sequence 1 (RW1):* First, we generated a random walk sequence of 60 s duration using a Markov chain with 76 discrete states and transition probabilities of stay (1/3), change to lower (1/3) or upper (1/3) adjacent states. To ensure that the this sequence formed a continuous multi-minute sequence so that participants did not notice begin or end of each segment, this initial 60 s sequence was concatenated with its temporally reversed version resulting in an RW1 sequence of 120 s duration. Each participant was presented with the RW1 sequence ~32 times during the experiment.

iii. *Random walk sequence 2 (RW2):* Likewise, we created a second random-walk sequence of 15 s duration using a Markov chain with only 38 possible states and transition probabilities similar to above. The 15 s sequence was concatenated with its temporally reversed version resulting in a 30 s sequence. The smoothness of this sequence segment was increased by filtering it (without phase shift) with a moving average of 250 ms. Each participant was presented with the RW2 sequence ~130 times.

In all sequences, the standard deviation spanned a range from 2-18°. The cloud's location mean was temporally independently resampled from five possible locations (-10°, -5°, 0°, 5°, 10°) at a stimulus onset asynchrony jittered between 1.4 and 2.8 s. In synchrony with the change in the cloud's location, the dots changed their colour and a sound was presented. The location of the sound was sampled from the two possible locations adjacent to the visual cloud's mean location. This ensured that the spatial discrepancy between the sound and visual cloud location was held constant at 5° to induce a strong common-source prior (Kording et al., 2007). The change in the dot's colour and the emission of sound occurred in synchrony to enhance audiovisual binding. The participants indicated the location of the sound by pressing one of 5 spatially corresponding buttons.

29 of the 56 participants participated in a sessions with sinusoidal and a session with RW1 sequence on different days. Eight additional participants only participated in the RW1 sequence. 18 independent participants participated in a session presenting the RW2 sequence. One participant completed all three sequences. Because the relative auditory weight (w_A, see below) was not significantly modulated by visual reliability of the current trial, we excluded five participants completing the Sin and RW1 sequence (i.e., inclusion criterion $p < 0.05$ in a linear regression of current visual reliability on w_A). Overall, we analyzed data from 25 participants for the sinusoidal, 33 participants for the RW1 and 19 participants for the RW2 sequence. We presented the sequences in 1676 trials, except in four sessions in which only 1128, 1143 or 1295 trials were presented. Before the experimental trials, participants practiced the auditory localization task in 25 unimodal auditory trials, 25 audiovisual congruent trials with a single dot as visual spatial cue and 75 trials with stimuli as in the main experiment.

Experimental setup

Audiovisual stimuli were presented using Psychtoolbox 3.09 (Brainard, 1997; Kleiner et al., 2007) (www.psychtoolbox.org) running under Matlab R2010b (MathWorks) on a Windows machine (Microsoft XP 2002 SP2). Auditory stimuli were presented at ~75 dB SPL using headphones (Sennheiser HD 555). Because visual stimuli required a large field of view, they were presented on a 30" LCD display (Dell UltraSharp 3007WFP). Participants were seated at a table in front of the screen in a darkened booth, resting their head on an adjustable chin rest. The viewing distance was 27.5 cm. This setup resulted in a visual field of approx. 100°. Participants gave response via a standard QWERTY keyboard. Participants used the buttons {i, 9, 0, -, =} with their right hand for localization responses.

Model free data analysis

To evaluate whether the relative influence of the auditory and the visual signals on the localization responses depended on the sequences' course of visual reliability, we first binned the localization responses into 20 bins according to each sequence (Fig. 6.2A-C). Using linear regression, we then predicted the localization responses by the auditory and visual signal location in each bin. We used the auditory ($ß_A$) and visual ($ß_v$) parameter estimates to compute the relative auditory influence as $w_A = ß_A / (ß_A + ß_v)$.

To determine whether the auditory weight w_A was a function of current and past visual reliability, we used a linear regression to predict w_A from the binned visual reliability and its temporal derivative in each participant. The temporal derivative captures influences of past reliability. To test the statistical significance of the influence of current and past visual reliability at the group level, we compared the parameters estimates of visual reliability and its temporal derivative against zero using one-sample t tests.

Further, we used the symmetry of the sequences (at period/2; cf. Fig. 6.2) to compare w_A from the first half of each sequence with the flipped w_A from the second half. This enabled us to estimate in each segment whether w_A differed given the same current visual reliability (cf. supplemental figure S6.1A), but different past visual reliabilities (Fig. 6.3A). To statistically evaluate the difference, we computed repeated measures ANOVAs on w_A with the factors bin (9 bins) and part of the sequence (first vs. second (flipped) half).

Computational models

To model the localization responses from the perspective of a Bayesian learner, we assumed that the participants used an internal generative model of the audiovisual signal (Fig. 6.1B). The experimenter presented the participant with an auditory signal $S_{A,t}$ at time t, together with a visual cloud of dots $S_{Vi,t}$. However, we assumed that the auditory signal that the brain has to process was corrupted by noise so that the internal auditory signal is $A_t \sim N(S_{A,t},\sigma_A)$, while the single visual dot (presented at high visual contrast) was uncorrupted, $V_{i,t} = S_{Vi,t}$. Due to our assumption of a common source, the location of the optimal estimate based on the auditory and visual signals was a weighted average of each of the signals

$$(1) \qquad \hat{S}_{AV,t} = \frac{\lambda_{V,t} V_t + \lambda_A A_t}{\lambda_t}$$

with the auditory reliability (i.e., the inverse of variance) $\lambda_A=1/\sigma_A^2$, visual reliability $\lambda_{V,t} = 1/\sigma_V^2$ and $\lambda_t = \lambda_A + \lambda_{V,t}$. So far this was the standard reliability-weighted cue combination (Ernst and Banks, 2002). However, the reliability of the visual signal had to be estimated as well. Thus, we allow for the reliability of V, $\lambda_{V,t}$, to be fluctuating on a fixed timescale (similar to a random walk), and the optimal behavior is thus to combine prior and current knowledge to estimate the distribution of $\lambda_{V,t}$. We specify $\lambda_{V,t}$ through a gamma distribution with parameters α and β, but where the parameters get updated in a Bayesian fashion after each new set of data points (visual dots) arrive:

$$(2) \qquad \alpha_{t,\text{posterior}} = \alpha_{t,\text{prior}} + \frac{n}{2}$$

$$(3) \qquad \beta_{t,\text{posterior}} = \beta_{t,\text{prior}} + \frac{(\sum_i (V_i - \bar{V}_t)^2) + \lambda_0\, n\, (\bar{V}_t - \mu_0)^2}{2(\lambda_0 + n)}$$

n = 20 is the number of visual dots and to have a small effect of the prior for the mean estimate, we set $\lambda_0 = 0.01$ and $\mu_0 = 0$. Based on this, the expected visual reliability was given as

(4) $$\hat{\lambda}_{V,t} = \frac{\alpha_{t,\text{posterior}}}{\text{ß}_{t,\text{posterior}}}$$

To model the change of visual reliability as a random-walk process, we use an approximation and modify the parameters of the visual reliability in between trials by the free parameter ϕ:

(5) $$\alpha_{t+1,\text{prior}} = \alpha_{t,\text{posterior}}\,\phi$$

(6) $$\text{ß}_{t+1,\text{prior}} = \text{ß}_{t,\text{posterior}}\,\phi$$

This has the effect of leaving the expectations of $\hat{\lambda}_{V,t}$ intact while increasing the uncertainty of the estimate of the variable (given as α_t/β_t^2).

As the internal variable for the auditory stimulus was random and not directly under the control of the experimenters, we generated 10,000 samples from the auditory likelihood $A_t \sim N(S_{A,t},\sigma_A)$, and for each value of A_t we calculated the optimal response (as above). These samples provided us with a histogram of possible responses according to the model, and thus the likelihood of the model given the participant responses.

As alternatives for learning of visual reliability, we introduced 3 different models: Model A1 assumes that reliability changes completely on a trial-to-trial basis and, thus, that there is no point in learning it, $\hat{\lambda}_{V,t} = 1/\sigma_{V_t}^2$. Model A2 assumes a simpler exponential discounting of the reliability so that

(7) $$\hat{\lambda}_{V,t} = 1/\sigma_{V,t}^2\,(1-\gamma) + \lambda_{V,t-1}\,\gamma$$

Model A3 assumes that the brain estimates the changes in physical variability of the visual cloud of dots ($\sigma_{V_t}^2$-$\sigma_{V_{t-1}}^2$) and extrapolates based on this,

(8) $$\hat{\lambda}_{V,t} = 1/\sigma_{V,t}^2 + \mathrm{d}\lambda_{V,t} \quad \text{where} \quad \mathrm{d}\lambda_{V,t} = \mathrm{d}\lambda_{V,t-1} + \theta\,(1/\sigma_{V,t}^2 - 1/\sigma_{V,t-1}^2 - \mathrm{d}\lambda_{V,t-1})$$

i.e. it updates a running estimate of the change in reliability.

Parameters for each model (σ_A and ϕ, γ or θ) were fit by minimizing the likelihood of the parameter on an individual participant basis, using MATLAB's fminsearch with multiple initial conditions to avoid local minima.

As an absolute measure of model performance, we computed the coefficient of determination (Nagelkerke, 1991) for the five candidate models. To do relative model comparison, we compared the four candidate models using the Bayesian Information Criterion (BIC) as an approximation to the model evidence (Raftery, 1995).

To compare the localization responses given by the participants and predicted by the Bayesian learner, we computed the auditory weight w_A from the Bayesian learner's predictions exactly as for the behavioral data. We then compared the model's w_A the from the first half of the sequences to the flipped w_A from the second half of the sequences (Fig. 6.3B).

From the Bayesian learner's ϕ parameter we computed the half-life of the influence of past reliability. The ϕ parameter can be interpreted as the fraction of reliablity information which is kept from past signals. Thus, the half life Λ of past visual influence is $\Lambda = \log(0.5) / \log(\phi) / 5\text{Hz}$. To test potential differences of Λ between the sequences, we computed a non-parametric permutations test (n = 5000 permutations) on the logit-transformed ϕ (i.e., rendering it a normal variable). To construct the permutation test, we used the F value from a one-way ANOVA with the between-subject factor sequence (sinusoid vs. RW1 vs. RW2) as the test statistic.

6.4 Results

To test whether human observers estimate sensory reliability only from current or, moreover, learn it from past signals, we used a multisensory ventriloquist paradigm. In this paradigm, human observers integrate audiovisual spatial signals weighted proportional to their dynamically varying reliability (Battaglia et al., 2003; Alais and Burr, 2004). Thus, auditory signals attain a large relative weight in case of low and a small relative weight in case of high visual reliability. A small auditory weight shifts the perceived auditory location towards the visual location which is also perceived in the ventriloquist illusion (Radeau and Bertelson, 1977). Thus, the key question was whether the relative auditory weight depended on the reliability of the current or, moreover, past visual signals.

In the ventriloquist paradigm, we presented human participants with audiovisual signals randomly sampled from five possible locations (Fig. 6.1A). The visual signals consisted of clouds of dots which were presented at a rate of 5 Hz. Crucially, the clouds' variance (i.e., the inverse of their reliability) changed in periodic sequences according to a sinusoid (n = 25; period = 30 s), a random walk (RW1; n = 33; period = 120 s) or a smoothed random walk (RW2; n = 19; period = 30 s) (Fig. 6.2). In synchrony with the change in the cloud's mean location, the dots changed their colour and a slightly offset spatial sound (± 5° discrepancy) was presented. The task of the participants was to localize the sounds.

We used the participants' sound localizations to determine a relative auditory weight w_A for each of 20 segments of the variance sequences (Fig. 6.2). Using linear regression in each segment, we predicted the participants' sound localizations by the auditory and visual signal locations and computed w_A as the relative auditory parameters estimate (i.e., $w_A = ß_A / (ß_A + ß_V)$). Thus, w_A varies between one, pure auditory influence, and zero, pure visual influence. As predicted by reliability-weighted integration, we found for all three sequences that the auditory weight increased linearly (i.e., the visual influence decreased) if the current variance of the cloud of dots increased (p < 0.001 for all three sequences; one-sample t test against zero). Crucially, the auditory weight depended also on the change of variance between a segment and its precursor (p = 0.001 for sinusoid, p = 0.014 for RW1 and p = 0.028 for RW2). This result revealed that the participants took the variance of the current and, moreover, of past visual signals into account while weighting the audiovisual signals.

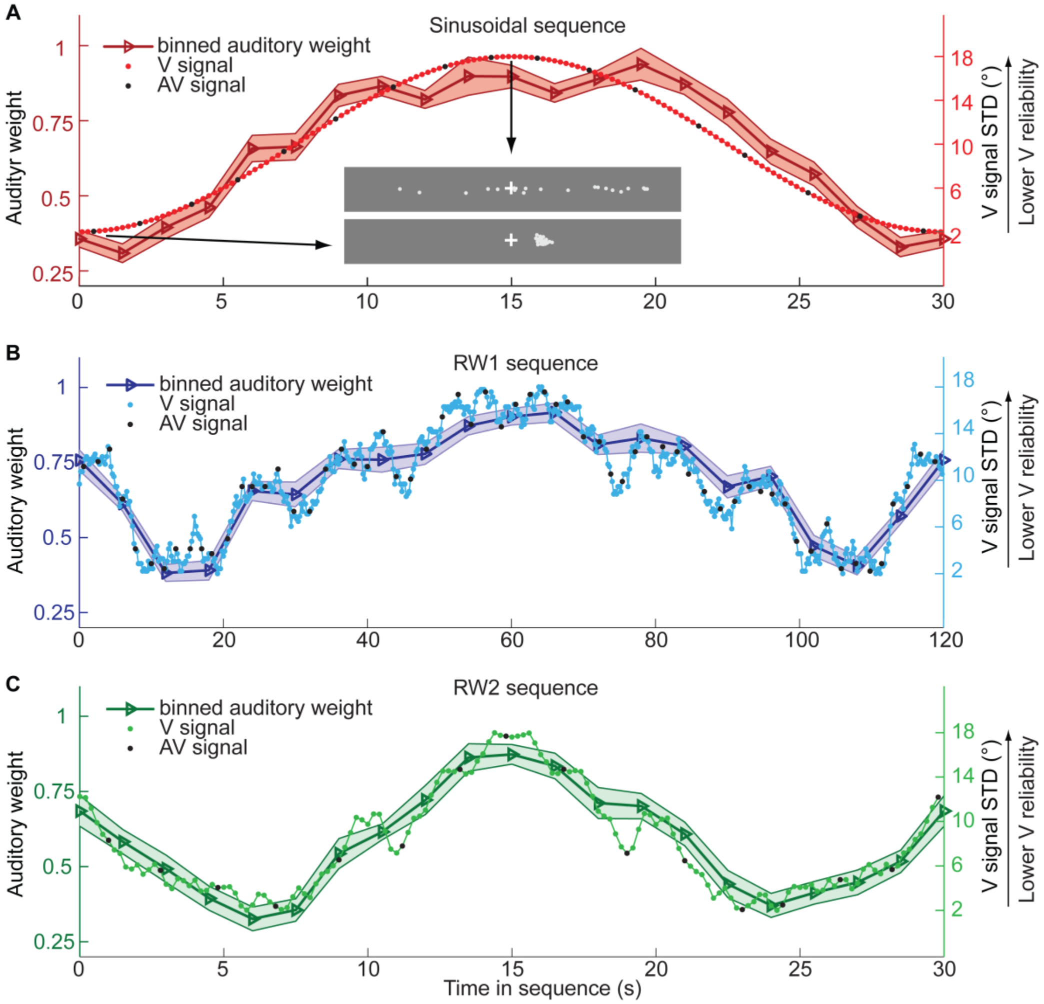


Figure 6.2. Time course of auditory weights in the three variance sequences. Binned (n = 20 bins) relative auditory weight (mean across participants ± SEM, left ordinate) as a function of the time in the sequence manipulating the variance of the visual (V) signal (i.e., the inverse of visual reliability, right ordinate). The relative auditory weight varies between one (i.e., pure auditory influence on the localization responses) and zero (i.e., pure visual influence). Visual variance was manipulated by (**A**) a sinusoid (period 30s, N = 25), (**B**) a random walk (RW1, period 120s, N = 33) and (**C**) a smoothed random walk (RW2, period 30s, N = 19). The sequence of visual signals was presented at 5 Hz while audiovisual (AV) signals (black dots) were interspersed with a temporal jitter. For illustration, the cloud of dots in case of the lowest (i.e., V signal STD = 2°) and the highest (i.e., V signal STD = 18°) visual variance are shown in (A).

To confirm this impression, we used the symmetry of the variance sequences to compare w_A from the first half with the flipped w_A from the second half of the sequences (Fig. 6.3A). This enabled us to directly compare w_A in each segment given the same current (cf. supplemental Fig. 6.1A), but different past variances of the cloud of dots. If participants reached a segment from a high level of visual variance, they gave the auditory signals a larger weight than reaching the same segment from a low level. Thus, we found a main effect of the factor sequence part (first vs. second half) in a repeated measures ANOVA for the sinusoid ($F_{1,\,24}$ = 12.162, p = 0.002, partial η^2 = 0.336) and the RW1 sequence, ($F_{1,\,32}$ = 14.129, p < 0.001). Further, we found an interaction effect of the factor sequence part and the segment for the RW2 sequence ($F_{4.6,\,82.9}$ = 3.385, p = 0.010) due to its non-monotonous course. Hence, the analysis again confirmed that participants used the history of visual variance to weight the audiovisual signals. However, the analyses could not reveal which specific strategy the participants used to estimate variance from its history.

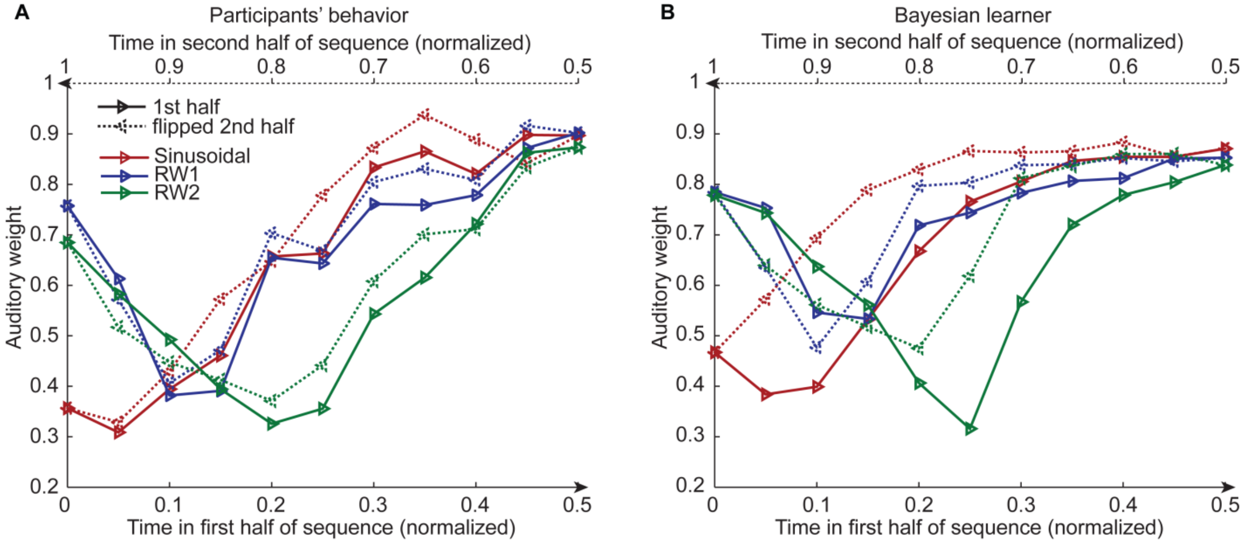


Figure 6.3. Comparison between the relative auditory weights from the participants' behavior and the Bayesian learner in the first and the second half of the variance sequences. (**A**) The relative auditory weights (mean across participants) are plotted as a function of the normalized time in the three sequences manipulating the variance of the visual cloud of dots. The first (lower abscissa) and the second—flipped—half (upper abscissa) of the auditory weights are plotted separately. (**B**) The relative auditory weights predicted by the Bayesian learner plotted in the same way as in (A).

Therefore, we tested whether the participants estimated visual variance like a Bayesian learner who optimally estimates (posterior) variance by updating the prior estimate of visual variance obtained from past visual signals with the variance estimate from the current signal (i.e., the likelihood) (Fig. 6.1B). Further, we compared the Bayesian learner with a learner who computes variance by exponentially discounting past variance and a learner who combines current variance with the variance expected from a linear extrapolation from past trials. The Bayesian learner's localization responses were very similar to the participants' localization responses (explained variance $R^2 = 66.8 \pm 2.6$ % (mean ± SEM) for the sinusoid, $R^2 = 66.5 \pm 2.4$ % for RW1 and $R^2 = 70.0 \pm 2.8$ % for RW2). Similar to the participants' relative auditory weights, the Bayesian learner's relative auditory weights depended on the visual variance of a given as well as the previous segment of a sequence (Fig. 6.3B; see supplemental Fig. S6.1B-C for the auditory weights of the remaining models). Moreover, the Bayesian learner outperformed the exponential discounting and extrapolation learners as well as a model which estimated reliability only from the current signal in nearly all participants (in 24 of the 25 participants for the sinusoid, 30/33 for RW1 and 19/19 for RW2; for model comparison details see tab. 6.1). Next, we inferred the half-life of the influence of past variance from a parameter of the Bayesian learner. We found that on average the participants included variance information from signals several seconds ago (2.4s for Sin, 5.6s for RW1 and 1.8s for RW2; no significant difference between sequences, $p = 0.470$ in permutation test).

Table 6.1. Model parameters and fit indices (mean ± SEM) for the four candidate models in the three sequences of visual variance.

Sequence	Model	σ_A	median Φ,γ,θ	R^2	BIC	pW	PP	EP
Sin	Bayesian learner	6.6 + 0.7	0.93	66.8 + 2.6	0	0.96	0.864	1
	Non-learner	4.0 + 0.2	-	51.1 + 7.1	440 + 79	0	0.034	0
	Exponential discounting	3.9 + 0.2	0.33	51.1 + 7.1	448 + 79	0	0.035	0
	Extrapolation	4.2 + 0.2	0.40	53.7 + 6.2	391 + 71	0.04	0.067	0
RW1	Bayesian learner	6.6 + 0.4	0.97	66.5 + 2.4	0	0.909	0.839	1
	Non-learner	4.4 + 0.1	-	57.0 + 4.6	48 + 299	0.091	0.106	0
	Exponential discounting	4.4 + 0.2	0.38	54.7 + 4.5	350 + 45	0	0.027	0
	Extrapolation	4.5 + 0.2	0.25	55.4 + 4.3	335 + 42	0	0.027	0
RW2	Bayesian learner	6.4 + 0.6	0.93	70.0 + 2.8	0	1	0.869	1
	Non-learner	4.1 + 0.2	-	56.7 + 6.9	71 + 398	0	0.045	0
	Exponential discounting	4.1 + 0.2	0.74	56.7 + 6.9	372 + 75	0	0.043	0
	Extrapolation	4.2 + 0.2	0.15	57.4 + 6.5	360 + 71	0	0.043	0

$\sigma_{A= \text{auditory}}$ variance; Φ = update parameter of the Bayesian learner; γ = discounting parameter of the exponential discounting model; θ = parameter of the extrapolation model; R^2 = coefficient of determination; BIC = Bayesian information criterion relative to best model (smaller = better) ; pW = proportion of participants in which model was better than any other model according to BIC; PP = posterior probability of model; EP = exceedance probability; (PP and EP computed from random-effects model comparison as implemented in SPM8, cf. Stephan 2009).

To sum up, the participants estimated sensory reliability by optimally combining current and prior reliability learned from past signals over several seconds. Thus, the participants estimated sensory reliability consistent with a Bayesian learner.

6.5 Discussion

From a Bayesian perspective, optimal percepts of physical quantities result from combining prior knowledge with new evidence provided by signals. Consistent with such a Bayesian strategy, we demonstrated that human observers learn posterior visual reliability by updating prior reliability from past with incoming reliability information from current signals. Perceptually, the observers estimated the location of audiovisual signals by weighting the signals proportional to the learned posterior visual reliability. Thus, the influence of past visual reliability on the perceived signal location dated back to visual signals several seconds ago.

Our results add an important aspect to the Bayesian perspective on perception (Yuille and Buelthoff, 1996; Knill and Pouget, 2004): To our knowledge, the current study is the first demonstration that human observers do not only learn the reliability of the prior (Kording et al., 2004; Kording and Wolpert, 2004; Berniker et al., 2010), but they also learn the evidence's reliability using Bayesian inference. Thus, Bayesian inference might fundamentally link sensory learning and perception (Fiser et al., 2010): On the one hand, human observers use Bayesian inference to learn the posterior sensory reliability by combining prior and current reliability information. On the other hand, they use Bayesian inference to perceive the actual physical quantity (i.e., the signal location) by weighting the signals proportional to the estimated posterior reliability. Even though we have only shown Bayesian learning of sensory reliability in a multisensory ventriloquist paradigm, we expect that such a Bayesian process might be fundamental when observers combine different kinds of unisensory (Jacobs, 1999; Knill and Saunders, 2003), multisensory (Ernst and Banks, 2002; Battaglia et al., 2003; Alais and Burr, 2004) and sensorimotor (Kording et al., 2004; Kording and Wolpert, 2004) signals.

Further, our findings suggest an extension of the theory of probabilistic population codes (Ma et al., 2006). For the gain of the neurons' population response parametrically represents a signal's reliability, the gain might be modulated by prior reliability. Alternatively, in a sampling-based neuronal representation of reliability, the neurons' activity encodes samples of a signal which are collected over time (Fiser et al., 2010). Sensory reliability is estimated from the samples' variability and, therefore, naturally learned over successive signals.

In conclusion, our results reveal that perception and sensory learning are two inextricable mechanisms by which the brain models the environment's current and past statistical properties. Both are governed by the same underlying principle — Bayesian inference.

6.6 Acknowledgments

This study was funded by the Max Planck Society and the European Research Council.

6.7 References

Alais D, Burr D (2004) The ventriloquist effect results from near-optimal bimodal integration. Curr Biol 14:257-262.

Algazi VR, Duda RO, Thompson DM, Avendano C (2001) The cipic hrtf database. In: Applications of Signal Processing to Audio and Acoustics, 2001 IEEE Workshop on the, pp 99-102: IEEE.

Battaglia PW, Jacobs RA, Aslin RN (2003) Bayesian integration of visual and auditory signals for spatial localization. J Opt Soc Am A Opt Image Sci Vis 20:1391-1397.

Berniker M, Voss M, Kording K (2010) Learning priors for Bayesian computations in the nervous system. PLoS One 5.

Brainard DH (1997) The psychophysics toolbox. Spatial vision 10:433-436.

Ernst MO, Banks MS (2002) Humans integrate visual and haptic information in a statistically optimal fashion. Nature 415:429-433.

Faisal AA, Selen LP, Wolpert DM (2008) Noise in the nervous system. Nat Rev Neurosci 9:292-303.

Fiser J, Berkes P, Orban G, Lengyel M (2010) Statistically optimal perception and learning: from behavior to neural representations. Trends Cogn Sci 14:119-130.

Jacobs RA (1999) Optimal integration of texture and motion cues to depth. Vision Res 39:3621-3629.

Kleiner M, Brainard D, Pelli D, Ingling A, Murray R, Broussard C (2007) What's new in Psychtoolbox-3. Perception 36:1.1-16.

Knill DC, Saunders JA (2003) Do humans optimally integrate stereo and texture information for judgments of surface slant? Vision Res 43:2539-2558.

Knill DC, Pouget A (2004) The Bayesian brain: the role of uncertainty in neural coding and computation. Trends Neurosci 27:712-719.

Kording KP, Wolpert DM (2004) Bayesian integration in sensorimotor learning. Nature 427:244-247.

Kording KP, Ku SP, Wolpert DM (2004) Bayesian integration in force estimation. J Neurophysiol 92:3161-3165.

Kording KP, Beierholm U, Ma WJ, Quartz S, Tenenbaum JB, Shams L (2007) Causal inference in multisensory perception. PLoS One 2:e943.

Ma WJ, Beck JM, Latham PE, Pouget A (2006) Bayesian inference with probabilistic population codes. Nat Neurosci 9:1432-1438.

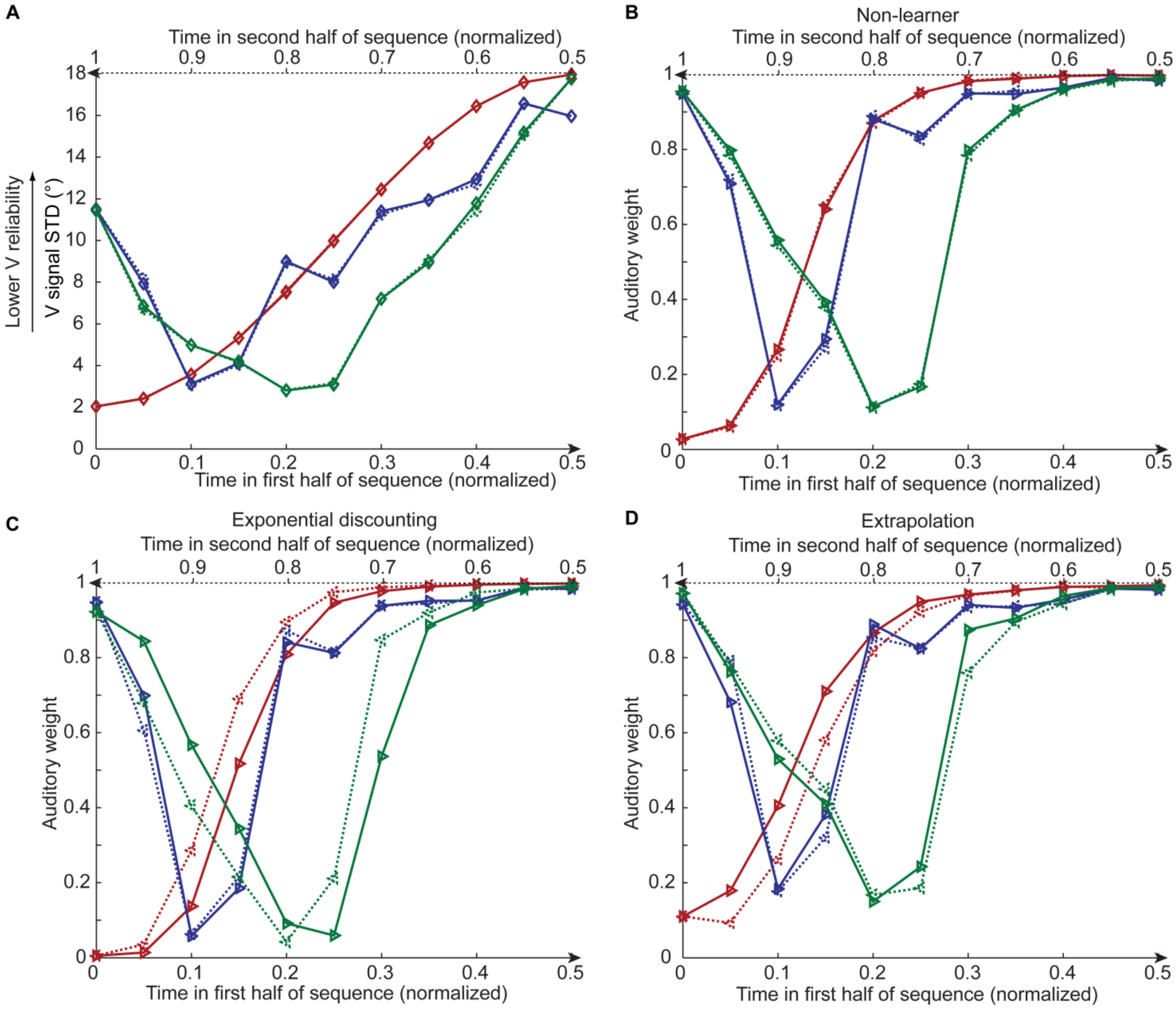
Nagelkerke NJ (1991) A note on a general definition of the coefficient of determination. Biometrika 78:691-692.

Radeau M, Bertelson P (1977) Adaptation to auditory-visual discordance and ventriloquism in semirealistic situations. Perception & Psychophysics 22:137-146.

Raftery AE (1995) Bayesian model selection in social research. Sociol Methodol 25:111-163.

Yuille AL, Buelthoff HH (1996) Bayesian decision theory and psychophysics. New York: Cambridge University Press.

6.8 Supplemental results



Supplemental figure S6.1. **Binned visual-variance sequences and auditory weights predicted by the alternative non-Bayesian models in the first and second half of the sequences. (A)** Comparison of visual (V) variance (i.e., STD of the cloud of dots) of the first and the second half of the sequences. Visual variance is binned into 20 segments just as the relative auditory weight w_A in Fig. 6.3A. Binning does not create any history effects. **(B, C, D)** The relative auditory weights (mean across participants) predicted by the alternative models are plotted as a function of normalized time in the three sequences manipulating the variance of the visual signal. The first (lower abscissa) and the second—flipped—half (upper abscissa) of the auditory weights are plotted separately. **(B)** Non-learner (model A1). **(C)** Exponential discounting (model A2). **(D)** Extrapolation (model A3).

7 The invisible ventriloquist

7.1 Abstract

Information integration across the senses is fundamental for effective interactions with our environment. A controversial question is whether multisensory integration is automatic or depends on perceptual awareness. Combining the spatial ventriloquist illusion and continuous flash suppression (dCFS), we investigated whether unconscious visual signals can influence conscious spatial perception of sounds. Importantly, dCFS obliterated visual awareness only in a fraction of trials allowing us to compare spatial ventriloquism for physically identical flashes that were visible or invisible. Our results show a stronger ventriloquist effect for visible than invisible flashes. Nevertheless, invisible flashes elicited a robust ventriloquist effect, even when participants were not better than chance on visual flash localization. These findings demonstrate that unconscious signals in one sensory modality can alter conscious perception in another sensory modality. They suggest that audiovisual signals can be integrated into spatial representations at least to some extent prior to perceptual awareness.

7.2 Introduction

Information integration is critical for effective interactions with our natural environment. To form a coherent and more reliable percept, the brain needs to integrate signals from multiple senses. It remains controversial, to what extent multisensory integration is automatic or dependent on higher cognitive processes such as attention or awareness (Talsma et al., 2010).

Accumulating evidence suggests that audiovisual integration depends on attention and awareness. Most prominently, the McGurk illusion falters under high attentional demands (Alsius, Navarra, Campbell, & Soto-Faraco, 2005). Likewise, the McGurk illusion is abolished when the visual facial movements are obliterated from awareness in the context of flash suppression (Palmer and Ramsey, 2012) or bistable perception (Munhall, ten Hove, Brammer, & Pare, 2009). These findings converge with the idea that consciousness enables the convergence and integration of information and processes within a global work space (Tononi and Edelman, 1998; Dehaene and Naccache, 2001; Baars, 2005). Yet, all previous studies have focused on the McGurk illusion, which illustrates integration of higher order phonological information (i.e., visemes and phonemes) during speech processing. Thus, it remains unclear whether consciousness is a general prerequisite for multisensory integration. Given accumulating evidence that multisensory integration emerges already at the primary cortical level (Foxe et al., 2000; Molholm et al., 2002), a critical question is whether low-level spatiotemporal information can be integrated automatically in the absence of attention or awareness.

Evidence for 'automatic' integration of spatial information comes predominantly from the ventriloquist illusion that emerges when sensory signals are artificially brought into spatial conflict (Radeau and Bertelson, 1977; Bertelson and Radeau, 1981). In spatial ventriloquism the perceived location of one sensory input (e.g., auditory) is shifted towards the location of a temporally correlated but spatially displaced input of another sensory modality (e.g., visual) and vice versa depending on the relative sensory reliabilities (Alais and Burr, 2004). Critically, ventriloquism has been observed even when decisional biases and response strategies are carefully controlled (Vroomen and de Gelder, 2004). Moreover, it has been shown to be unaffected by endogenous or exogenous spatial attention (Bertelson et al., 2000a; Vroomen et al., 2001). In fact, ventriloquism is thought to facilitate and hence occur prior to spatial attentional selection (Driver, 1996). Likewise, ventriloquism was not influenced by modality-specific attention, i.e. whether participants focused on a particular sensory modality (Vroomen and de Gelder, 2004).

Collectively, this body of research suggests that spatial ventriloquism emerges at the sensory processing level largely unaffected by attentional or decisional control. Therefore, one may ask whether it emerges even prior to or in the absence of participants' awareness. Initial tentative evidence from patients with spatial hemineglect suggests that audiovisual spatial ventriloquism persists for visual signals that participants are not aware of (Bertelson et al., 2000b). Yet, these results need to be interpreted with caution because the ventriloquist effect was reported as significant only for visual signals in patients' neglected, but not in their intact hemifield. Furthermore, this study characterized the ventriloquist effect only for unaware but not for aware visual signals in patients' neglected hemifield. Therefore, it could not directly compare the effects of visible and invisible signals to formally quantify the contributions of awareness to audiovisual spatial integration.

To investigate whether integration of auditory and visual signals into multisensory spatial representations relies on perceptual awareness, the present study combined spatial ventriloquism with dynamic continuous flash suppression (dCFS) (Tsuchiya and Koch, 2005; Maruya et al., 2008). Dynamic CFS suppresses participants' awareness of monocularly viewed events by simultaneously presenting rapidly changing motion grating masks to the other eye (Maruya et al., 2008). Using dCFS, we presented participants' suppressed eye with a brief visual flash to their left or right hemifield. In synchrony with the flash, a brief beep was played in the centre, left or right hemifield. Critically, we selected the saliency of the visual flash, such that the dynamic continuous flash suppression obliterated visual awareness only in a fraction of trials. This allowed us to compare spatial ventriloquism for physically identical flashes that do or do not enter participant's awareness.

7.3 Materials and methods

Participants

After giving informed consent, 32 healthy young adults (19 females, 30 right-handed, mean age: 23.5 years, standard deviation: 3.53, range: 18-38) with normal or corrected-to-normal vision, participated in this study. One subject was excluded, because of non-compliance; he provided random responses during the auditory localization task as indicated by an approximately zero correlation between true auditory stimulus locations and auditory localization responses. The study was approved by the local ethics review board of the University of Tübingen.

Stimuli and apparatus

Participants sat in a dimly lit room in front of a computer monitor at a viewing distance of 1 m. They viewed one half of the monitor with each eye using a custom-built mirror stereoscope. Visual stimuli were composed of targets and masks that were presented on a grey, uniform background with a mean luminance of 15.5 cd/m^2. One eye viewed the target stimuli, i.e. two grey discs (Ø 0.29°, mean luminance: 25.4 cd/m^2), located 5.72° visual angle to the left and right of a grey fixation dot. On each trial, either the left or the right target's colour changed to white (mean luminance: 224.2 cd/m^2) for a duration of 100 ms. This change in brightness will be referred to as 'flash'.

To suppress the flash's perceptual visibility, two dynamic Mondrians (Ø 2°) were shown to the other eye (Maruya et al., 2008). To match the target's location, the Mondrians' were also centred 5.72° to the left and right of the fixation dot. Each Mondrian consisted of sinusoidal gratings (Ø 0.57°) which changed their colour and position randomly at a frequency of 10 Hz. Each grating's texture was shifted every 16.6 ms to generate apparent motion. Visual stimuli were presented foveally, contained a fixation spot and were framed by a grey, isoluminant square aperture of 8.58° x 13.69° in diameter to aid binocular fusion.

Auditory stimuli were pure tones with a carrier frequency of 1 kHz and a duration of 100 ms. They were presented via six external speakers, placed above and below the monitor. Upper and lower speakers were aligned vertically and located centrally, 2.3° to the left and 2.3° to the right of the monitor's centre. Speakers' location was chosen by trading off physical alignment of visual and auditory stimulus locations and sound localization performance. At a distance of 2.3°, mean sound localization accuracy amounted to ~70% (see below).

Psychophysical stimuli were generated and presented on a PC running Windows XP using the Psychtoolbox version 3 (Brainard, 1997; Kleiner et al., 2007) running on Matlab 7 (Mathworks, Nantucket, Massachusetts). Visual stimuli were presented dichoptically using a gamma-corrected 30” LCD monitor with a resolution of 2560 x 1600 pixels at a frame rate of 60Hz (GeForce 8600GT graphics card). Auditory stimuli were digitized at a sampling rate of 44.8 kHz via a M-Audio Delta 1010LT sound card and presented at an maximal amplitude of 73 dB sound pressure level. Exact audiovisual onset timing was confirmed by recording visual and auditory signals concurrently with a photo-diode and a microphone.

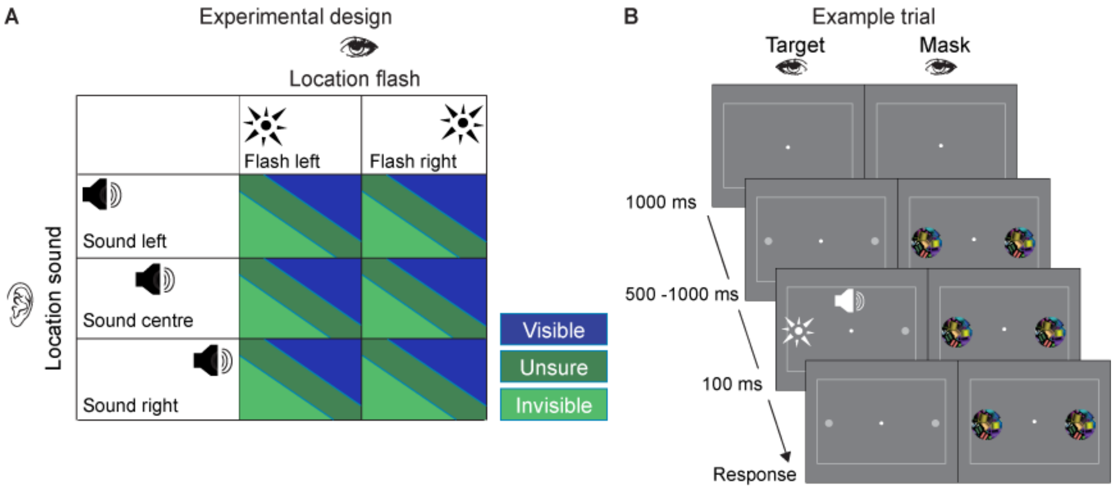


Figure 7.1. Experimental paradigm and procedure. (A) Experimental design: 2 x 3 x 3 factorial design with the factors: 1. Flash location (left, right) 2. Sound location (left, centre, right) 3. Visibility (Visible, Unsure, Invisible). **(B)** Example trial and procedure of dynamic flash suppression

Experimental Design

In a spatial ventriloquist paradigm, participants were presented with an auditory beep emanating from one of three potential locations: left, centre, right. In synchrony with the beep, one eye was presented with a brief flash either in participants’ left or right hemifield under dynamic continuous flash suppression (Maruya et al., 2008). Hence, the 2 x 3 x 3 factorial design manipulated 1. ‘flash location’ (2 levels: left flash and right flash), 2. sound location (3 levels: left sound, central sound and right sound) and 3. flash visibility (3 levels: visible, unsure, invisible) (Figure 7.1).

Each trial started with the presentation of the fixation dot for a duration of 1000 ms. Next, participants’ one eye was presented with two grey discs, located 5.72° visual angle to the left and right of a grey fixation dot. Participants’ awareness of these discs was suppressed by showing dynamic Mondrians to the other eye (i.e., dynamic continuous flash suppression). The Mondrian masks and the discs were presented on the screen until participants had responded to all questions. The assignment of eyes was changed after each

trial, to enhance suppression. After a random interval of 500-1000 ms either the left or the right disc 'flashed', i.e. changed its luminance for a duration of 100 ms. In synchrony with the flash, an auditory beep was played from one of three potential locations.
On each trial, participants reported the location of the beep (left, centre, right) and rated the visibility of the flash (visible, unsure, invisible). This visibility judgment provided a 'subjective awareness criterion'. Critically, the flash was visible only in a fraction of trials allowing us to quantify the effect of awareness on multisensory integration by comparing spatial ventriloquism for physically identical flashes that were visible or invisible. The 'unsure' response option was primarily included to encourage participants to categorize trials as invisible. Participants responded by pressing one of three buttons on a keyboard. The button assignment was counterbalanced across participants.

In addition, on 22.2% of the trials, the so-called catch trials, participants were also asked to locate the flash (left vs. right discrimination; in addition to visibility judgment and sound localization). This allowed us to assess the spatial information that is available for visual spatial localization during visible, unsure and invisible trials and select participants that were not better than chance when locating flashes that they judged as invisible (i.e., the so-called chance performers). The latter allowed us to investigate the influence of flashes on sound localization, when they were invisible and unaware in an objective sense (i.e., objective awareness criterion).

Prior to the main experiment, participants were familiarized with stimuli and task. First, they completed 2-3 sessions of sound localization (% correct day1: 69.9% (std.: 16.8); % correct day2: 74.7% (std.: 14.7), % correct both days: 72.3% (std.14.9)). Next, there were two short practice sessions of the ventriloquist paradigm. During the main experiment participants completed a total of 24 experimental sessions distributed over two successive days, resulting in a total of 1296 trials (i.e., 216 trials per condition).

Analysis

For data analysis, participants' perceived auditory location was coded as -1 for left, 0 for centre and 1 for right across trials. For each participant, we estimated the crossmodal bias $(A_{responded} - A_{loc})/(V_{loc} - A_{loc})$ as an index of the spatial ventriloquist effect with $A_{responded}$ = participant's auditory location response, V_{loc} = the location of the visual signal and A_{loc} = the location of the auditory signal. To account for subject-specific spatial response biases and limited response options, we adjusted A_{loc} and V_{loc} using a linear regression approach across all congruent non-catch trials irrespective of visibility level. Therefore, we linearly regressed participants' localization responses against the true signal location in the congruent trials and inserting the predicted auditory and visual locations as A_{loc} and V_{loc} in the crossmodal bias equation. This adjustment recovered the 'true' ventriloquist effect reliably in particular for small visual influence as in our current study (cf. simulations performed in supplemental fig. S7.1A). As an alternative measure of crossmodal bias, we regressed the auditory and visual locations against the auditory location responses and computed the crossmodal bias as the relative visual weight. This approach yielded very similar results (cf. supplemental fig. S7.1B).

7.4 Results

Spatial ventriloquism for visible, unsure and invisible trials

First we computed the ventriloquist effect (i.e., crossmodal bias) separately for each visibility level. To maximize the power of the analysis, we included all subjects that had at least one incongruent trial for a particular visibility level (n.b. this analysis approach naturally results in different number of subjects being included for different visibility levels). As shown in figure 7.2A, the ventriloquist effect was much stronger for visible than unsure and invisible trials. Thus, a repeated measures ANOVA with the factor visibility (visible, unsure, invisible) revealed a significant main effect of visibility ($F_{1.1,30.3} = 14.709$; $p < 0.001$; $\eta^2 = 0.353$; Greenhouse-Geisser-corrected).

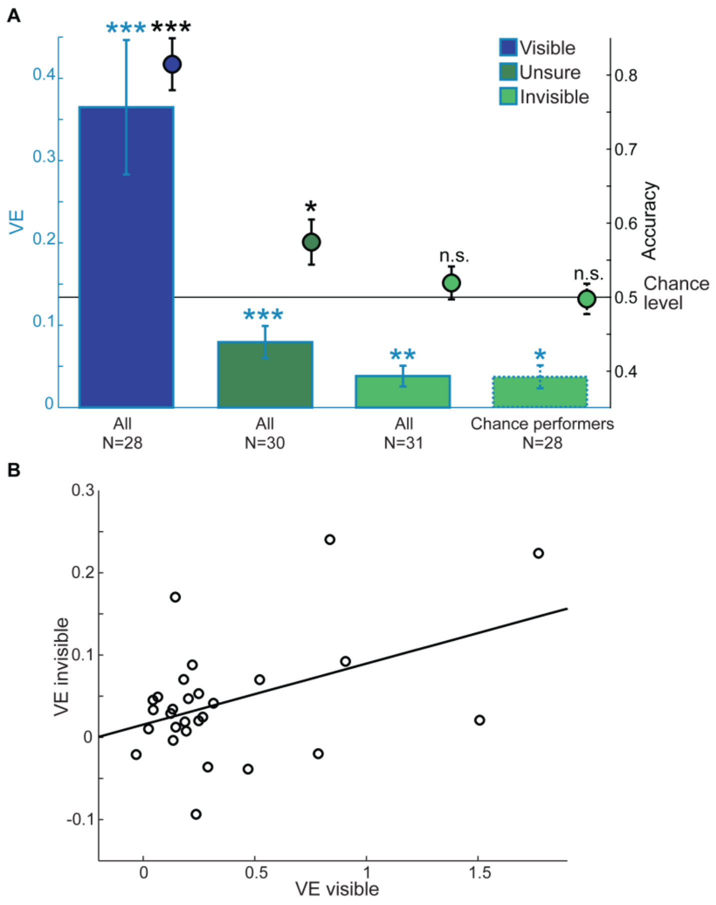


Figure 7.2. The ventriloquist effect (VE), flash localization performance and the correlation between the ventriloquist effect for visible versus invisible flashes. (A) Bar plots show the ventriloquist effect (= crossmodal bias; left ordinate) for trials where participants judged the visibility of the flash as 'visible', 'unsure' or 'invisible'. We either included all participants that had trials at the respective visibility level (= All) or only those that performed at chance on flash localization on trials where they responded 'invisible' (= chance performers). The markers show the flash localization performance at the group level (across included subjects mean ± SEM; right ordinate). For invisible trials, participants were not better than chance at the group level when including all subjects or only the subjects that were at chance based on individual binomial testing. **(B)** Scatter plot depicting the correlation between the ventriloquist effect for visible and invisible flashes over subjects.

Nevertheless, a highly significant ventriloquist effect was observed when testing separately for visible trials (t_{27} = 4.466, p < 0.001), unsure trials (t_{29} = 4.039, p < 0.001) and invisible trials (t_{30} = 3.025, p = 0.005). Importantly, the ventriloquist effect for invisible trials emerged even though at the group level participants performed at chance on the flash localization task (Fig. 7.2A; flash localization accuracy for invisible trials (across-subjects mean ± SEM): 0.518 ± 0.022; t-test against 0.5 chance performance: t_{30} = 0.788, p = 0.437).

Correlation between visible and invisible ventriloquist effects

Next, we investigated whether the ventriloquist for visible and invisible trials were correlated over subjects. In other words, we determined whether participants that show a strong (resp. weak) ventriloquist effect for invisible trials also exhibit a strong (resp. weak) ventriloquist effect for visible trials. As shown in figure 7.2B, the ventriloquist effects for visible and invisible trials were significantly correlated over subjects (r = 0.444, p = 0.018, n = 28, i.e., including all subjects with visible and invisible trials). This correlation provides initial evidence that the neural mechanisms and circuitries underlying the ventriloquist effects in the presence and absence of awareness may be at least partly overlapping.

Spatial ventriloquism for invisible trials restricted to chance performers on flash localization

In the first analysis, we demonstrated a ventriloquist effect for invisible trials, when subjects were at chance at the group level when locating invisible flashes. Thus, we used a more stringent criterion of perceptual awareness by including only those subjects that were individually not better than chance when locating an ‘invisible’ flash during the catch trials using a binomial test (i.e., objective awareness criterion). The individual chance performance constraint reduced the number of subjects that could be included in the analysis (n = 28). Nevertheless, despite the reduced number of subjects, we still observed a significant ventriloquist effect for invisible trials (Fig. 7.2A; t_{27} = 2.630, p = 0.014). Moreover, as shown in figure 7.2A, the size of the ventriloquist effect is similar when including all subjects or only the chance performers. Importantly, the flash localization accuracy (i.e., across subjects mean) at the group level is again not significantly better than chance but nearly equal to 50% (flash localization accuracy: 0.498 ± 0.021 (across subjects mean ± SEM; t test against 0.5 chance performance: t_{27} = -0.118, p = 0.9080).

Regression analysis: VE prediction by flash localization accuracy

Next, we investigated whether the ventriloquist effect was predicted by participants’ accuracy to localize the flash (n.b. this analysis could be performed only on participants that had data in both catch and non-catch trials for the respective visibility levels). As shown in figure 7.3A, we observed a significantly positive regression slope for visible trials (all participants: $t_{23} = 2.623$, $p = 0.009$; chance performers: $t_{20} = 2.518$, $p = 0.010$). This positive prediction of the ventriloquist effect for visible flashes by visual localization accuracy is consistent with models of Bayes-optimal integration where the sensory weight increases with the reliability of the signal (Ernst and Banks, 2002; Alais and Burr, 2004). If participants have access to highly reliable visual information as indicated by their flash localization accuracy, they show a strong ventriloquist effect.

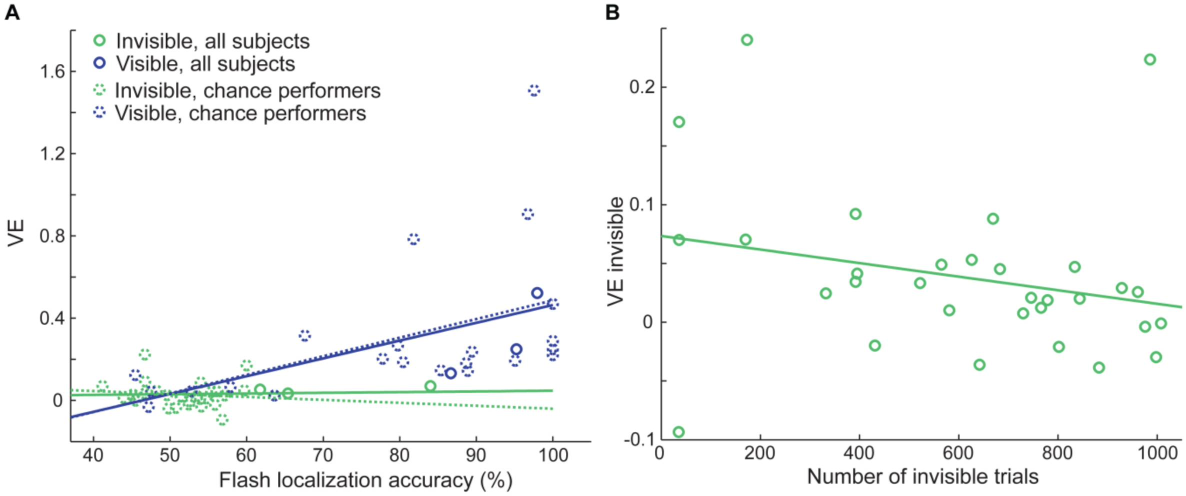


Figure 7.3. Relation of ventriloquist effect (VE), flash localization accuracy and number of trials with invisible flashes. (A) The scatter plots depict the regression of the ventriloquist effect against localization accuracy of the visual flash separately for visible (blue) and invisible (green) trials based on all subjects (solid) or only chance performers (dashed). The ordinate represents the ventriloquist effect and the abscissa represents visual localization accuracy (i.e., % correct). **(B)** The scatter plot depicts the regression of the ventriloquist effect for invisible flashes against the number of trials with invisible flashes.

Critically, however, for invisible flashes, the ventriloquist effect did not significantly depend on the accuracy with which participants were able to locate the flash. This was true when all participants were included ($t_{28} = 0.236$, $p = 0.408$; one outlier subject with an accuracy of zero in a single invisible catch trial was excluded; yet, when this outlier subject was included, we observed even a negative regression slope) or when only the chance performers were included ($t_{25} = -0.556$, $p = 0.291$). These results indicate that visual representations may influence sound processing without being available for visual localization tasks or accessible to perceptual awareness. They also provide further support that the ventriloquist effect for invisible trials does not arise predominantly from participants with higher flash localization accuracies.

Regression analysis: VE prediction by the number of invisible trials

One may hypothesize that the ventriloquist effect arises predominantly in subjects that set a very high criterion for judging flashes as visible and may therefore still have visual information available on trials that are judged as invisible. To address this potential concern, we investigated whether the ventriloquist effect was predicted by the number of invisible trials per subject. However, contrary to this conjecture, we observed a regression slope that was not significantly different from zero ($p > 0.05$; Fig. 7.3B). Thus, this regression analysis provides additional corroborative evidence that the ventriloquist effect for invisible trials is not driven predominantly by participants that set a high visibility criterion and hence may still have visual information available when they judge a flash as invisible.

7.5 Discussion

Using continuous flash suppression and spatial ventriloquism, we demonstrate that unconscious signals in the visual modality influence how humans construct their auditory perceptual world. In particular, we have shown that invisible flashes alter the perceived location of concurrent sounds. These results suggest that auditory and visual inputs are integrated into coherent spatial representations at least to some extent prior to perceptual awareness.

Accumulating evidence has shown that audiovisual integration of speech signals is abolished when visual facial movements are rendered unconscious via multistable perception or flash suppression (Munhall et al., 2009; Palmer and Ramsey, 2012) highlighting the role of perceptual awareness in multisensory integration. This raises the question whether consciousness is a generic prerequisite for multisensory integration and is also required for low-level spatial integration as indexed by the ventriloquist effect.

Our findings demonstrate that spatial ventriloquism is profoundly modulated by the visibility of the flash. While a strong ventriloquist effect was observed for visible trials, it was attenuated, when the visual flash was not consciously perceived. Nevertheless, a robust ventriloquist effect persisted for invisible trials, even when participants showed chance performance on flash localization. Moreover, in additional regression analyses we demonstrated that the ventriloquist effect for invisible trials was not significantly predicted by participant's flash localization accuracy or the number of invisible flashes. These findings further corroborate that the ventriloquist effect for invisible trials is not driven by participants that can still access visual information despite judging the stimulus as invisible. Collectively, our results demonstrate that 'invisible' flashes that evade participants' awareness influence where we perceive sounds that we are aware of.

At least two distinct neural circuitries may mediate the influence of these 'invisible' flashes on sound localization during continuous flash suppression. First, an invisible flash may interact with auditory signals via subcortical mechanisms such as the colliculo-pulvinar pathway (Wallace et al., 1993; Hackett et al., 2007; Cappe et al., 2009b; Cappe et al., 2009a) that has previously been implicated in mediating activations along the dorsal stream into the intraparietal sulcus under CFS (Fang and He, 2005). Second, it may modulate sound processing via sparse direct connectivity between primary auditory and visual areas (Falchier et al., 2002; Cappe and Barone, 2005).

The ventriloquist effect may be smaller for invisible than visible flashes, because ‘invisible’ flashes may evoke weaker activations than visible flashes already at the primary cortical level as a result of state-dependent effects or various sources of internal neural noise (Faisal et al., 2008). The level of neural activity then concurrently determines (i) whether the flash is able to enter perceptual awareness as well as (ii) the precision of the spatial representation and thereby strength of the ventriloquist effect (cf. Alais and Burr, 2004; Ma et al., 2006). Thus, visible flashes would induce a ventriloquist effect via the same neural circuitries as invisible flashes and induce a greater ventriloquist effect, as they induce higher neural activity and thus more precise spatial representations in visual cortices. In support of this ‘shared neural mechanism’ account, the size of the ventriloquist effect correlated significantly for ‘visible’ and ‘invisible’ flashes over subjects.

Alternatively, ‘visible’ flashes may induce a stronger ventriloquist effect by employing additional neural circuitries that are not engaged by weaker invisible flashes. This account dovetails nicely with current perspectives on the neural organization of multisensory integration. Specifically, auditory and visual information are thought to be integrated via multiple circuitries including subcortical mechanisms, direct connectivity between primary sensory areas and convergence in higher order association areas (Macaluso and Driver, 2005; Ghazanfar and Schroeder, 2006; Musacchia and Schroeder, 2009; Kayser et al., 2012). Moreover, it is well established that multisensory integration progressively increases along the cortical hierarchy with only about 15 % neurons showing multisensory properties in primary sensory areas (Bizley et al., 2007) and more than 50 % in classical association areas such as intraparietal or superior temporal sulci (Dahl et al., 2009). Thus, when a visual flash escapes the continuous flash suppression and enters participants’ awareness, a strong ventriloquist effect emerges most likely via integration in association areas such as intraparietal sulci (IPS) that contain exuberant multisensory neurons and may potentially amplify multisensory integration via feed-back loops with lower-level sensory areas. By contrast, when continuous flash suppression blocks neural activity at least to some extent from propagating into higher-order association areas, audiovisual integration is greatly attenuated or even abolished leading to a smaller ventriloquist effect mediated via an alternative neural circuit. Under this ‘multiple neural circuitries’ account, auditory and visual signals are integrated most likely at both pre- and post-aware processing stages potentially via partly distinct neural circuitries (e.g., direct connectivity vs. higher order association cortices).

In conclusion, to our knowledge our findings provide the first convincing demonstration that unconscious signals in one sensory modality can alter our conscious percept of signals in another sensory modality. These results suggest that low level sensory information can be integrated at least to some extent prior to perceptual awareness. Nevertheless, information integration as indexed by spatial ventriloquism was strongly amplified for conscious relative to unconscious visual signals. This raises the possibility that 'aware' visual signals may also engage multisensory integration mechanisms in higher-order association areas or other neural circuitries that are not engaged in the absence of perceptual awareness. Future studies using EEG and fMRI are needed to identify the neural systems that enable audiovisual integration in the presence and absence of awareness.

7.6 Acknowledgments

This research was funded by the Max Planck Society and the European Research Council. We thank Mario Kleiner for help with stimulus generation and building the experimental set up, Beatrix Barth and Natalie Christner for help with data collection, Joana Leitão for support and helpful discussions and Ulrik Beierholm for comments on a previous version of the manuscript.

7.7 References

Alais D, Burr D (2004) The ventriloquist effect results from near-optimal bimodal integration. Curr Biol 14:257-262.

Baars BJ (2005) Global workspace theory of consciousness: toward a cognitive neuroscience of human experience. Progress in brain research 150:45-53.

Bertelson P, Radeau M (1981) Cross-modal bias and perceptual fusion with auditory-visual spatial discordance. Perception & psychophysics 29:578-584.

Bertelson P, Vroomen J, de Gelder B, Driver J (2000a) The ventriloquist effect does not depend on the direction of deliberate visual attention. Percept Psychophys 62:321-332.

Bertelson P, Pavani F, Ladavas E, Vroomen J, de Gelder B (2000b) Ventriloquism in patients with unilateral visual neglect. Neuropsychologia 38:1634-1642.

Bizley JK, Nodal FR, Bajo VM, Nelken I, King AJ (2007) Physiological and anatomical evidence for multisensory interactions in auditory cortex. Cereb Cortex 17:2172-2189.

Brainard DH (1997) The Psychophysics Toolbox. Spatial vision 10:433-436.

Cappe C, Barone P (2005) Heteromodal connections supporting multisensory integration at low levels of cortical processing in the monkey. The European journal of neuroscience 22:2886-2902.

Cappe C, Rouiller EM, Barone P (2009a) Multisensory anatomical pathways. Hearing research 258:28-36.

Cappe C, Morel A, Barone P, Rouiller EM (2009b) The thalamocortical projection systems in primate: an anatomical support for multisensory and sensorimotor interplay. Cereb Cortex 19:2025-2037.

Dahl CD, Logothetis NK, Kayser C (2009) Spatial organization of multisensory responses in temporal association cortex. J Neurosci 29:11924-11932.

Dehaene S, Naccache L (2001) Towards a cognitive neuroscience of consciousness: basic evidence and a workspace framework. Cognition 79:1-37.

Driver J (1996) Enhancement of selective listening by illusory mislocation of speech sounds due to lip-reading. Nature 381:66-68.

Ernst MO, Banks MS (2002) Humans integrate visual and haptic information in a statistically optimal fashion. Nature 415:429-433.

Faisal AA, Selen LP, Wolpert DM (2008) Noise in the nervous system. Nat Rev Neurosci 9:292-303.

Falchier A, Clavagnier S, Barone P, Kennedy H (2002) Anatomical evidence of multimodal integration in primate striate cortex. J Neurosci 22:5749-5759.

Fang F, He S (2005) Cortical responses to invisible objects in the human dorsal and ventral pathways. Nature neuroscience 8:1380-1385.

Foxe JJ, Morocz IA, Murray MM, Higgins BA, Javitt DC, Schroeder CE (2000) Multisensory auditory-somatosensory interactions in early cortical processing revealed by high-density electrical mapping. Brain Res Cogn Brain Res 10:77-83.

Ghazanfar AA, Schroeder CE (2006) Is neocortex essentially multisensory? Trends Cogn Sci 10:278-285.

Hackett TA, De La Mothe LA, Ulbert I, Karmos G, Smiley J, Schroeder CE (2007) Multisensory convergence in auditory cortex, II. Thalamocortical connections of the caudal superior temporal plane. The Journal of comparative neurology 502:924-952.

Kayser C, Petkov CI, Remedios R, Logothetis NK (2012) Multisensory Influences on Auditory Processing: Perspectives from fMRI and Electrophysiology. In: The Neural Bases of Multisensory Processes (Murray MM, Wallace MT, eds). Boca Raton (FL).

Kleiner M, Brainard D, Pelli D (2007) What's new in Psychtoolbox-3? Perception 36.

Ma WJ, Beck JM, Latham PE, Pouget A (2006) Bayesian inference with probabilistic population codes. Nat Neurosci 9:1432-1438.

Macaluso E, Driver J (2005) Multisensory spatial interactions: a window onto functional integration in the human brain. Trends Neurosci 28:264-271.

Maruya K, Watanabe H, Watanabe M (2008) Adaptation to invisible motion results in low-level but not high-level aftereffects. Journal of vision 8:7 1-11.

Molholm S, Ritter W, Murray MM, Javitt DC, Schroeder CE, Foxe JJ (2002) Multisensory auditory-visual interactions during early sensory processing in humans: a high-density electrical mapping study. Brain Res Cogn Brain Res 14:115-128.

Munhall KG, ten Hove MW, Brammer M, Pare M (2009) Audiovisual integration of speech in a bistable illusion. Current biology : CB 19:735-739.

Musacchia G, Schroeder CE (2009) Neuronal mechanisms, response dynamics and perceptual functions of multisensory interactions in auditory cortex. Hearing research 258:72-79.

Palmer TD, Ramsey AK (2012) The function of consciousness in multisensory integration. Cognition 125:353-364.

Radeau M, Bertelson P (1977) Adaptation to auditory-visual discordance and ventriloquism in semirealistic situations. Attention, Perception, & Psychophysics 22:137-146.

Talsma D, Senkowski D, Soto-Faraco S, Woldorff MG (2010) The multifaceted interplay between attention and multisensory integration. Trends Cogn Sci 14:400-410.

Tononi G, Edelman GM (1998) Consciousness and complexity. Science 282:1846-1851.

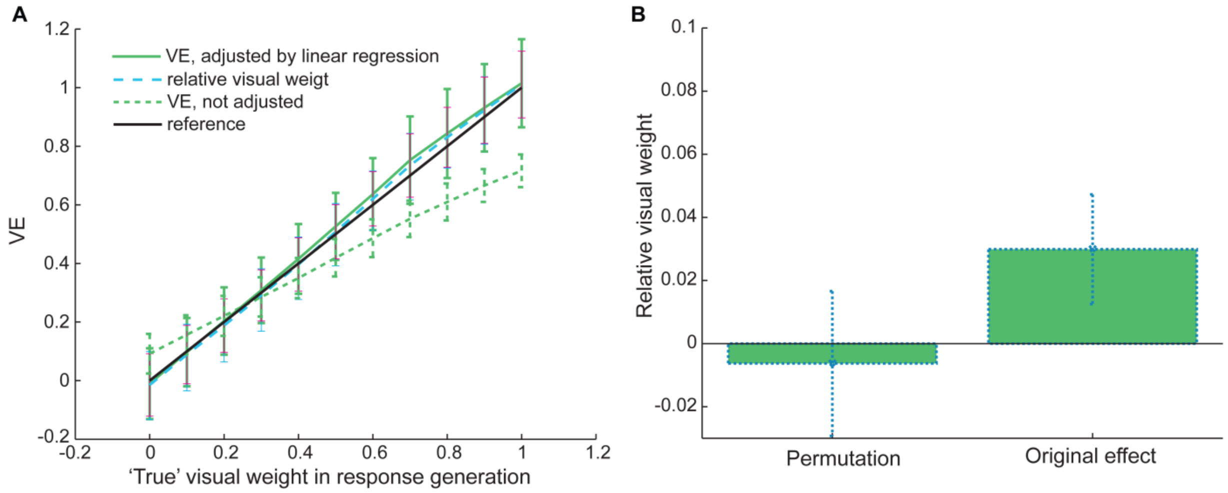
Tsuchiya N, Koch C (2005) Continuous flash suppression reduces negative afterimages. Nat Neurosci 8:1096-1101.

Vroomen J, de Gelder B (2004) Perceptual effects of cross-modal stimulation: Ventriloquism and the freezing phenomenon. In: The handbook of multisensory processes (Calvert G, Spence C, Stein BE, eds), pp 141 - 146. MA, USA: MIT Press.

Vroomen J, Bertelson P, de Gelder B (2001) The ventriloquist effect does not depend on the direction of automatic visual attention. Percept Psychophys 63:651-659.

Wallace MT, Meredith MA, Stein BE (1993) Converging influences from visual, auditory, and somatosensory cortices onto output neurons of the superior colliculus. Journal of neurophysiology 69:1797-1809.

7.8 Supplemental results



Supplemental figure S7.1. Simulation results showing validity of the adjusted ventriloquist effect (VE) as a measure of crossmodal bias. (A) To show the validity of the VE adjusted by linear regression as reported in the main text, we first simulated 'noisy auditory localization responses' according to $A_{response} = A_{loc}*ß_A + V_{loc}*ß_v + e$ with $A_{response}$ participants' auditory localization response, A_{loc}, V_{loc} the true auditory or visual locations and $ß_A$, $ß_v$ the relative auditory and visual weights (i.e., $ß_v = (1 - ß_A)$). The simulation used exactly the same parameters as our experiment (i.e., 2 visual and 3 auditory locations and three 3 response 'buttons' for the auditory localization responses). The figure shows the 'true/reference' $ß_v$, the naïve (unadjusted) VE, the VE adjusted by linear regression, and the relative visual weight ($ß_{relativeV}$, cf. (B)) (mean ± STD across 5000 simulations) as a function of $ß_v$. Only the adjusted VE and relative visual weight are unbiased. **(B)** Alternatively, the unbiased ventriloquist effect can be measured based on the relative visual weight that is computed as $ß_{relativeV} = ß_v / (ß_v + ß_A)$ using linear regression(i.e., $A_{response} = A_{loc}* ß_A + V_{loc}* ß_v + e$). While this ventriloquist index is less commonly used in the literature, it has the advantage that a permutation distribution under the null-hypothesis can be generated: Because the visual signal has no effect on auditory localization responses under the null hypothesis, the permutation distribution (mean ± STD across 5000 simulations) was generated by exchanging the left versus right label of the visual signal. As shown in (B), $ß_{relativeV}$ approximates zero for the permutation distribution, i.e. is unbiased under the null hypothesis. However, $ß_{relativeV}$ (mean ± SEM across subjects) is positive when computed for our original data from chance performers in trials of invisible flashes (cf. Fig. 7.2A). Here, $ß_{relativeV}$ is significantly greater than zero based on this non-parametric permutation testing ($p = 0.003$).