EJN European Journal of Neuroscience

FENS Federation of European Neuroscience Societies

European Journal of Neuroscience, Vol. 51, pp. 1161–1178, 2020 doi:10.1111/ejn.13813

Formal models and quantitative measures of multisensory integration: a selective overview

Hans Colonius[1,2,*] and Adele Diederich[2,3,*]
[1]Department of Psychology, Carl von Ossietzky Universität Oldenburg, Oldenburg 26111, Germany
[2]Department of Psychological Sciences, Purdue University, West Lafayette, IN, USA
[3]Life Sciences and Chemistry, Jacobs University Bremen, Bremen, Germany

Keywords: Bayesian causal inference, crossmodal recalibration, diffusion model, probability summation, race model, time window of integration

Abstract

Multisensory integration (MI) is defined as the neural process by which unisensory signals are combined to form a new product that is significantly different from the responses evoked by the modality-specific component stimuli. In recent years, MI research has seen exponential growth in the number of empirical and theoretical studies. This study presents a selective overview of formal modeling approaches to MI. Emphasis is on models and measures for behavioral paradigms, such as localization, judgment of temporal order or simultaneity, and reaction times, but some concepts for the modeling of single-cell spike rates are treated as well. We identify a number of essential concepts underlying most model classes, such as Bayesian causal inference, probability summation, coactivation, and time window of integration. Quantitative indexes for measuring and comparing the strength of MI across different paradigms are also discussed. Whereas progress over the last years is remarkable, we point out some strengths and weaknesses of the modeling approaches and discuss some obstacles toward a unified theory of MI.

Introduction

Investigating the processes involved in merging information from different sensory modalities has become a focus of research in many areas of anatomy, physiology, and behavior, as witnessed by an exponential increase in journal articles and documented in several recent handbooks (Calvert *et al.*, 2004; Naumer & Kayser, 2010; Bremner *et al.*, 2012; Murray & Wallace, 2012; Stein, 2012). Early studies by Todd (1912), measuring reaction time to stimuli from two or more sensory modalities presented both singly and together, are often seen as the beginnings of the scientific study of crossmodal interaction. Much later, more dramatic phenomena like the McGurk–McDonald effect (McGurk & MacDonald, 1976) in speech perception or, at the neurophysiological level, the identification of "multisensory" neurons in several brain areas (Meredith & Stein, 1983) have inspired a myriad of studies also including developmental aspects, brain disorders, and various applications like driver assistance systems.

In contrast to the huge body of empirically established phenomena, efforts to lay the theoretical foundations of multisensory integration and, in particular, to develop formalized quantitative models and measures have been somewhat less impressive. Over the years, a number of rather ubiquitous empirical "rules" characterizing multisensory processing have been identified, and certain quantitative measures of crossmodal effects are widely used to date. However, the role of these rules in current modeling approaches is not yet completely clarified. Given that the empirical database comprises several different levels of granularity, from the study of single neurons and neural populations to behavior with varying complexity and diverse methodologies, the prospects of a unified theory of multisensory processing appear somewhat poor, and its pursuit arguably may not be a promising research strategy at this time.

The focus of this review was on formalized quantitative modeling approaches with a focus on the underlying concepts. Obviously, this leaves out the bulk of important and valuable theoretical work on aspects that are not, or not yet, amenable to formalization. What then is the aim of developing quantitative models for multisensory integration? At its best, mathematical modeling in general serves to clarify which phenomena are to be explained; based on precise terms and concepts of a theory previously described only verbally; in particular, it may allow separating the essential assumptions of a theory from those that are merely auxiliary; moreover, it enables

Correspondence: Hans Colonius, as above.
E-mail: hans.colonius@uol.de

*H.C and A.D. contributed equally to this work.

Received 5 August 2017, revised 18 December 2017, accepted 20 December 2017

Edited by Dr. Sophie Molholm.

Reviewed by Benjamin Rowland, Wake Forest University, USA Luigi Acerbi, University of Geneva, Switzerland.

The associated peer review process communications can be found in the online version of this article.

precise quantitative predictions testable in subsequent experiments that lead to rejection or modification of the models.

Scrutinizing current modeling efforts for multisensory integration shows, however, that a realization of this program faces several challenges. First, a formal definition of "multisensory integration" and "multisensory processing" is still under debate, and consequently, alternative measures for quantifying the phenomenon exist (see below). Second, and perhaps most important, current models and measures are typically limited to a specific experimental paradigm and/or level of investigation like neuronal or psychophysical, as the often deplored "gap between spikes and behavior" of multisensory modeling testifies. Third, certain crossmodal effects may arguably not be amenable to formalization in principle, like some synesthetic experiences (Parise & Spence, 2009).

Measuring multisensory integration: from neurons to reaction times

Developing a measure of multisensory integration requires, above all, an agreement on what is going to be measured. In a recent review article in this journal (Stein *et al.*, 2010), sixteen authors, working in areas from basic neuroscience to psychology, identified a widespread semantic confusion among different investigative groups about several terms around "multisensory integration", with possibly severe consequences for scientific progress. In their glossary of terms, they distinguish properties of stimuli (modality-specific vs. crossmodal) from neural or behavioral properties (unisensory vs. multisensory). "Multisensory integration" is then (ibid, p. 1719) defined as

Multisensory integration *The neural process by which unisensory signals are combined to form a new product. It is operationally defined as a multisensory response (neural or behavioral) that is significantly different from the responses evoked by the modality-specific component stimuli.*

Moreover, "multisensory processing" should be interpreted as a "generic overarching term describing processing involving more than one sensory modality but not necessarily specifying the exact nature of the interaction between them". Thereby, paradigms like crossmodal matching, where stimuli from different modalities are compared to estimate their perceptual equivalence, and synesthesia are included, as well.

A key feature of the above definition is the "significant difference" between the crossmodal response and the modality-specific responses, as it directly leads to the original quantitative index of multisensory integration for the number of spikes emitted in a fixed time interval after stimulation (Meredith & Stein, 1983; Stein & Meredith, 1993).

Definition

$$\mathrm{CRE} = \frac{\mathrm{CM} - \mathrm{SM}_{\max}}{\mathrm{SM}_{\max}} \times 100, \qquad (1)$$

where CM is the mean (absolute) number of spikes in response to the crossmodal stimulus and $\mathrm{SM}_{\max}$ is the mean (absolute) number of spikes to the most effective modality-specific component stimulus. Thus, CRE expresses crossmodal enhancement as a proportion of the strongest unisensory response. For simplicity, we define CRE and its variations below at the population level. In applications, these would be estimated by sample values.

Some modifications of CRE have been proposed as well (Perrault *et al.*, 2005). Prominently, in the "additive model", term $\mathrm{SM}_{\max}$ in Eqn (1) is replaced by the sum of the unisensory responses (Populin & Yin, 2002). This additive version has raised some controversy because, under some modeling assumptions, an additive combination of crossmodal inputs yields a prediction of optimal multisensory integration (Pouget *et al.*, 2002). Thus, observing that a neural circuit is actually engaged in optimal multisensory enhancement but does not achieve "superadditivity" would lead one to conclude that no multisensory integration has taken place. Similarly, any crossmodal response larger than the largest unisensory response but smaller than the sum might be "misinterpreted" as response depression (Stein *et al.*, 2009). Nevertheless, the additive version is often used, in addition to the original CRE, to measure the amount of "superadditivity" (e.g., Miller *et al.*, 2017). At this point, it is important to point out that our discussion here is necessarily incomplete given that the important case of multisensory inhibition has been omitted, due to limits of space.

Modifying Definition 1 to become applicable to measures other than spike numbers is straightforward. For reaction times, multisensory integration typically manifests itself by speeded responses to crossmodal stimuli (Todd, 1912; Diederich, 1992a). Then, CRE (Eqn 1) becomes (Diederich & Colonius, 2004a),

$$\mathrm{CRE}_{\mathrm{RT}} = \frac{\min\{\mathbf{E}\mathrm{RT}_V, \mathbf{E}\mathrm{RT}_{A+\tau}\} - \mathbf{E}\mathrm{RT}_{VA}}{\min\{\mathbf{E}\mathrm{RT}_V, \mathbf{E}\mathrm{RT}_{A+\tau}\}} \times 100, \qquad (2)$$

where $\mathbf{E}\mathrm{RT}_{VA}$ is the mean RT to the crossmodal stimulus (taken as combination of visual and acoustic stimuli here) and $\min\{\mathbf{E}\mathrm{RT}_V, \mathbf{E}\mathrm{RT}_{A+\tau}\}$ is the faster of the unisensory mean RTs to the visual stimulus and the acoustic stimulus presented at stimulus-onset asynchrony (SOA) τ, for $-\infty < \tau < +\infty$. Thus, $\mathrm{CRE}_{\mathrm{RT}}$ expresses multisensory enhancement as a proportional reduction in the faster unisensory response by the crossmodal response. For example, $\mathrm{CRE}_{\mathrm{RT}} = 10$ means that mean response time to the visual–auditory stimulus is 10% faster than the faster of the expected response times to unimodal visual and auditory stimuli. Finally, for detection probabilities an analogous measure is

$$\mathrm{CRE}_{\mathrm{DP}} = \frac{p_{VA} - \max\{p_V, p_A\}}{\max\{p_V, p_A\}}, \qquad (3)$$

where p_V, p_A, p_{VA} refer to the detection probabilities (relative frequencies, in applications) under the different conditions.

While CRE and its variations have obvious value as simple descriptive measurement tools, we have recently taken issue with their appropriateness as universal operational definition for the amount of multisensory enhancement (Colonius & Diederich 2017). The starting point is the observation that being responsive to multiple sensory modalities does not guarantee that a neuron has actually engaged in integrating its multiple sensory inputs, rather than simply responding to the most effective stimulus in a given trial, that is, to the stimulus eliciting the strongest response (Stein *et al.*, 2009). Assuming random variation of the responses, such a mechanism may produce significant levels of CRE by a mechanism known as probability summation, although it would not be considered "true" multisensory integration as it does not actually combine the activities elicited by the modality-specific stimulus components ("coactivation"). Interestingly, it can be shown that the expected value (in a sample: the average) of the random number of spikes elicited by a

crossmodal stimulus combination will be maximal when probability summation operates under (maximal) negative statistical dependency, that is, when large responses to one stimulus component co-occur with small responses to the other component, and vice versa. The measure defined next takes this maximum as point of reference:

$$\mathrm{CRE}^{-} = \frac{\mathrm{CM} - \mathrm{SM}^{-}}{\mathrm{SM}^{-}} \times 100. \quad (4)$$

Here, SM^{-} replaces $\mathrm{SM}_{\max}$ in CRE; it refers to the mean maximally achievable by combining the unisensory data under a probability summation rule with negative dependence. CRE^{-} is, in general, sensitive to the entire distribution of spikes (in particular, the variance) and also more conservative than CRE in never predicting more multisensory integration. In a sample of spike data, we recently showed that this new measure tends to reduce the proportion of neurons so far labeled as "multisensory" using a statistical criterion (see Colonius & Diederich, 2017, for further details).

Moreover, it has been shown that, in the context of reaction time data, the logic underlying measure CRE^{-} is identical to one being traditionally used to assess the amount of integration by testing the race model inequality (see Eqn (26) in Section "Models for reaction time paradigms" below). Alternative versions for CRE_{RT} and CRE_{DP}, based on probability summation analogous to measure CRE^{-} in Eqn (4), have also been developed (Colonius & Diederich, 2006; Colonius, 2015).

Note that, while CRE^{-} takes a probability summation mechanism as reference model, it is not postulated that the system actually operates that way: it only serves as a well-defined benchmark against which to gauge the crossmodal responses. Developing a specific model of multisensory integration requires to identify further factors that may affect measures of multisensory integration. For example, in a recent study (Miller *et al.*, 2015), Miller and colleagues found that the relative difference between the response magnitude (i.e., mean number of spikes per trial) to the visual (V) and auditory (A) stimuli, called unisensory imbalance (UI),

$$\mathrm{UI} = \frac{V - A}{V + A}, \quad (5)$$

plays a decisive role in predicting the strength of multisensory integration. Specifically, their findings suggest that "reducing the effectiveness of one unisensory component in a pair will cause it be integrated more efficaciously when the stronger stimulus is "early" rather than "late"." (ibid, p. 5216).

In summary, research on (i) how to define the amount of multisensory integration across different paradigms and response modes and (ii) what these measures depend upon is ongoing, at least as long as multisensory modeling has not yet finished (Beauchamp, 2005; Perrault *et al.*, 2005; Stein *et al.*, 2009).

Modeling multisensory integration: What needs to be modeled?

A defining feature of multisensory integration is the distinction between a modality-specific context where stimuli of a single modality are presented, and a crossmodal context where stimuli from two or more modalities are presented. For concreteness, we refer to $\mathcal{V}, \mathcal{A}, \mathcal{T}$ as the unimodal context where visual, auditory, or tactile stimuli are presented within a spatiotemporal frame, respectively. Similarly, $\mathcal{VA}$ denotes a bimodal (visual–auditory) context, etc.

Any multisensory integration model needs to specify how responses to signals from the unisensory contexts, for example $\mathcal{V}$ and $\mathcal{A}$, are related to responses to signals from the crossmodal context, for example $\mathcal{VA}$. The nature of this relationship may take many different forms, depending on how stimuli are defined in space and time, the experimental paradigm, and the level of description (single neuron level, neuronal subpopulations, neuro-imaging, behavioral). Moreover, the response in a crossmodal context $\mathcal{VA}$ may not be reducible to some deterministic function of the responses in $\mathcal{V}$ and $\mathcal{A}$ (e.g., the McGurk–McDonald effect where the multisensory percept is different from both the visual and the auditory unisensory percept).

Most existing models endeavor to account for some of the ubiquitous empirical "rules" or principles referred to above. Perhaps most important, the "spatiotemporal rule" holds that stimuli presented in spatiotemporal proximity have a higher probability of being integrated (the "spatiotemporal window" of integration, see Meredith, 2002). Second, the "inverse-effectiveness rule" predicts that the amount of integration increases when the intensity level of the unimodal components is lowered. Third, according to the "reliability rule", the sensory modality with the higher reliability (e.g., lower signal-to-noise ratio) is weighted more heavily when integrating unisensory signals. None of these "rules" may apply in a given situation but some, like the "spatiotemporal rule", are often found in both behavioral and neural data.

This raises a number of issues that need to be addressed. First, what are the costs and benefits of multisensory integration? Is it mandatory or are there instances where information from one modality is simply ignored? Second, if integration does occur, what are the mechanisms controlling this process? In particular, does integration follow some rules of optimality? It turns out that there is a huge array of multisensory modeling approaches with different underlying theoretical frameworks.

There are many possible criteria to classify multisensory modeling approaches. Apart from requiring well-defined concepts and internal consistency, a multisensory integration model should allow one to derive predictions for empirical data under specified conditions, even if only for a very limited set of experimental paradigms. We have chosen to assemble the modeling approaches along a number of important paradigms, both behavioral and neural, but with an emphasis on the former.

Models for psychophysical estimation and judgment paradigms: Cue combination

In all psychophysical paradigms considered here, information from more than one sensory modality is available to the observer. Typical tasks are, for example, estimating the location of a sound source in a visual–acoustic environment, comparing the size of two visual–haptic stimuli presented in two consecutive temporal intervals, or deciding whether two stimuli presented in different modalities occur simultaneously or not. While, in principle, all methods and models developed within psychophysics could be relevant to analyze data obtained under these paradigms, the focus on multisensory modeling, as followed here, is on how information from the different modalities is combined. Many psychophysical tasks require that an observer estimate object properties such as location, orientation, or shape. Many of these properties, like brightness, are intramodal, that is, specific to a single sensory modality, while others can be assessed by drawing on information, or cues, from two or more senses simultaneously, like location or shape, these latter being referred to as supramodal. Models of cue combination based on decision-theoretic concepts, including multidimensional signal

detection theory, have been around in psychology at least since the 70s and 80s of the last century (Green & Swets, 1966; Shaw, 1982). In a similar vein, introducing Bayesian concepts into perceptual modeling has a long tradition (see Knill & Richards, 1996, for a collection of papers reflecting the state of research in the 90s). Instead of tracing the history of these modeling approaches, we limit our presentation here to some central ideas.

The standard "Gaussian" cue combination model

Let s be the physical property of some stimulus, for example, the size of an object to be estimated by an observer via two sensory modalities, for example, vision and touch (haptic). Following signal detection theory (SDT), repeated presentation of the stimulus gives rise to a probability distribution of (subjective) perceptual values in each sensory modality, S_V and S_H for visual and haptic stimuli. In the Gaussian version, S_V and S_H are assumed to be independent, normally distributed random variables with shared expected value

$$\mathbf{E}S_V = \mathbf{E}S_H = s$$

and possibly different variances σ_V^2 and σ_H^2, respectively. We wish to estimate s given S_V and S_H using a weighted linear combination

$$w_V S_V + w_H S_H$$

with non-negative weights w_V and w_H and $w_V + w_H = 1$. Obviously, this is an unbiased estimator of s,

$$\mathbf{E}[w_V S_V + w_H S_H] = w_V \mathbf{E}S_V + w_H \mathbf{E}S_H = s,$$

with variance

$$\mathbf{V}[w_V S_V + w_H S_H] = w_V^2 \sigma_V^2 + w_H^2 \sigma_H^2.$$

Estimates of the variances can be obtained by different psychophysical paradigms. For example, Ernst & Banks (2002) fit Gaussian psychometric functions to visual–haptic discrimination frequencies of standard vs. comparison stimuli under different noise conditions.

The model holds that the weights w_V and w_H are determined in such a way that the reliability of the weighted linear combination is maximal, with reliability being defined as the reciprocal of the variance:

$$r_i = \sigma_i^{-2} \quad \text{for} \quad i = V, H.$$

The following optimal solution is adapted from Oruç *et al.* (2003):

Proposition

The variance of $w_V S_V + w_H S_H$, with $w_V + w_H = 1$, is minimized when the weights are

$$w_V = \frac{\sigma_V^{-2}}{\sigma_V^{-2} + \sigma_H^{-2}} = \frac{r_V}{r_V + r_H} \text{ and } w_H = \frac{\sigma_H^{-2}}{\sigma_V^{-2} + \sigma_H^{-2}} = \frac{r_H}{r_V + r_H}$$

yielding

$$\mathbf{V}[w_V S_V + w_H S_H] = \left(\frac{1}{\sigma_V^2} + \frac{1}{\sigma_H^2}\right)^{-1} = \frac{\sigma_V^2 \sigma_H^2}{\sigma_V^2 + \sigma_H^2} \leq \min\{\sigma_V^2, \sigma_H^2\}.$$

Moreover, the minimum-variance linear combination reaches maximum reliability, r_W, that is,

$$r_W = (w_V^2 r_V^{-1} + w_H^2 r_H^{-1}) = r_V + r_H. \tag{6}$$

Remarkably, the proof does not require an assumption about the distribution of S_V and S_H, except for existence of the variance. Unfortunately, this is often not recognized when the linear "Gaussian cue combination" model is probed on a set of data. On the other hand, when nonlinear cue combinations are possible, the Gaussian assumption is sufficient to prove the minimum-variance-unbiased property of the weighted linear combination. As in this case the solution is also the maximum-likelihood estimate, the model is often called MLE model.

Proof

The proof uses a special case of the Cauchy–Schwarz inequality (see e.g., Mitrinović, 1970, p. 30) that follows from simple algebraic transformations:

Lemma

For any real numbers a_V, a_H, b_V, b_H

$$(a_V b_V + a_H b_H)^2 \leq (a_V^2 + a_H^2)(b_V^2 + b_H^2)$$

with equality if and only if (i) $b_V = b_H = 0$ or (ii) a_V is proportional to b_V and a_H is proportional to b_H.

To show Proposition 1, we set

$$a_i = w_i \sigma_i \quad \text{and} \quad b_i = \sigma_i^{-1} \quad \text{for} \quad i = V, H.$$

From the lemma,

$$(w_V + w_H)^2 \leq (w_V^2 \sigma_V^2 + w_H^2 \sigma_H^2)\left(\frac{1}{\sigma_V^2} + \frac{1}{\sigma_H^2}\right). \tag{7}$$

The first term on the right-hand side is the variance to be minimized, and the other terms on that side do not depend on choice of the weights. A minimum is obviously achieved when inequality (7) is an equality. As the $b_i = \sigma_i^{-1}$ cannot be zero, this occurs when there is a constant of proportionality, c, such that, for $i = V, H$,

$$a_i \equiv w_i \sigma_i = c b_i = c \sigma_i^{-1}.$$

Thus, $w_i = c/\sigma_i^2$ and $w_H + w_V = 1$ implies

$$c = \frac{1}{\frac{1}{\sigma_V^2} + \frac{1}{\sigma_H^2}}$$

This implies the statement of Proposition 1 immediately.

The case of correlated cue distributions and negative weights is discussed in Oruç *et al.* (2003) as well. In particular, the authors show that observers may employ negative weights in some cue combination tasks, using a less reliable cue to cancel "noise" from a more reliable cue when the cues are highly correlated. Parise *et al.* (2012) have probed the model in a localization task with visual, auditory, and combined stimulus streams. The improvement in precision for combined relative to unimodal targets was statistically optimal only when audiovisual signals were correlated. Thus, their subjects used the similarity in the temporal structure of multisensory signals to solve the correspondence problem, hence inferring causation from correlation.

For a comprehensive discussion of relaxing various assumptions of cue combination models see Ernst (2012). Despite its simplicity, the basic optimal cue integration model has been reasonably successful across many psychophysical tasks, with some notable exceptions (for a recent review, see Fetsch *et al.*, 2013).

Bayesian causal inference models

Up to now, cue combination has been considered as mandatory leading to a more reliable estimate of the crossmodal stimulus. Consider the case where a large discrepancy occurs between the values of, say, S_V and S_A, the subjective values triggered by a visual–auditory stimulus complex. The observer cannot tell whether this is due (i) to random noise in the processing of neuronal signals or (ii) to some systematic difference between the signals. In case (i), by integrating the two estimates the perceptual system could average out the influence of such noise, as discussed in the previous model. In case (ii), the difference may be the result of S_V and S_A not being generated by a common event or object, or of an additive bias in one or both variables such that, for example, $\mathbf{E}S_A = s + b_A$ or $\mathbf{E}S_V = s + b_V$ (see Odegaard *et al.*, 2015, for a study of uni- and bimodal biases in audiovisual perception of space).

Facing this inherent uncertainty about the source of discrepancy, the causal inference model (Körding *et al.*, 2007; Shams & Beierholm, 2010) distinguishes two possible causes: both S_V and S_A may have been the result of a common audiovisual event/object or they may have been generated by different events/objects. The observer uses two pieces of information: One piece is the likelihood of the observed values of S_V and S_A under the common event or under different events, and the other piece is the prior probability of whether a common event (or source) exists. The model has been formulated in the context of an visual–auditory spatial localization task, but it applies more generally.

With indicator variable C for a common event $(C = 1)$ or two different events $(C = 2)$, let $\Pr(C = 1) = 1 - \Pr(C = 2)$. If $C = 1$, the source locations are the same, $s = s_V = s_A$, and the observer draws a position s from a prior distribution $\mathcal{N}(0, \sigma_p)$, where $\mathcal{N}(\mu, \sigma)$ stands for a normal distribution with mean μ and standard deviation σ. This expresses the prior belief that it is more likely that a stimulus is centrally located (0°) than in the periphery. If $C = 2$, positions s_V and s_A are each drawn independently from $\mathcal{N}(0, \sigma_p)$. In keeping with the notation in Wozny *et al.* (2010), x_V, x_A stand for the perceptual values of the event, that is, realizations of random variables S_V, S_A with probability distribution $p(x_V, x_A)$, and they are also drawn independently from $\mathcal{N}(s_V, \sigma_V)$ and $\mathcal{N}(s_A, \sigma_A)$, respectively. Thus, the perceptual noise is assumed to be independent across modalities.

The posterior probability for C based on the sensory evidence is obtained via Bayes's rule,

$$p(C \mid x_V, x_A) = \frac{p(x_V, x_A \mid C)p(C)}{p(x_V, x_A)}. \tag{8}$$

Going over to the ratio and taking logarithms, we obtain

$$\log\frac{\Pr(C = 1 \mid x_V, x_A)}{\Pr(C = 2 \mid x_V, x_A)} = \log\frac{p(x_V, x_A \mid C = 1)}{p(x_V, x_A \mid C = 2)} + \log\frac{\Pr(C = 1)}{\Pr(C = 2)}. \tag{9}$$

The first term on the right is the log-likelihood ratio with the numerator obtained by marginalizing across all values of position s

$$p(x_V, x_A \mid C = 1) = \int p(x_V, x_A \mid s)p(s)\mathrm{d}s = \int p(x_V \mid s)p(x_A \mid s)p(s)\mathrm{d}s. \tag{10}$$

For the denominator, marginalization is across both possible positions,

$$\begin{aligned} p(x_V, x_A \mid C = 2) &= \int\int p(x_V, x_A \mid s_V, s_A)p(s_V, s_A)\mathrm{d}s_V\mathrm{d}s_A \\ &= \left(\int p(x_V \mid s_V)p(s_V)\mathrm{d}s_V\right)\left(\int p(x_A \mid s_A)p(s_A)\mathrm{d}s_A\right). \end{aligned} \tag{11}$$

Explicit expressions for the likelihood functions under the above assumption of independent Gaussians are found in Körding *et al.* (2007). Under these assumptions, the optimal estimates for the common cause model are

$$\hat{s}_{A,C=1} = \hat{s}_{V,C=1} = \frac{\dfrac{x_A}{\sigma_A^2} + \dfrac{x_V}{\sigma_V^2} + \dfrac{\mu_p}{\sigma_p^2}}{\dfrac{1}{\sigma_A^2} + \dfrac{1}{\sigma_V^2} + \dfrac{1}{\sigma_p^2}}; \tag{12}$$

and for the independent $(C = 2)$ model,

$$\hat{s}_{A,C=2} = \frac{\dfrac{x_A}{\sigma_A^2} + \dfrac{\mu_p}{\sigma_p^2}}{\dfrac{1}{\sigma_A^2} + \dfrac{1}{\sigma_p^2}} \quad \text{and} \quad \hat{s}_{V,C=2} = \frac{\dfrac{x_V}{\sigma_V^2} + \dfrac{\mu_p}{\sigma_p^2}}{\dfrac{1}{\sigma_V^2} + \dfrac{1}{\sigma_p^2}}. \tag{13}$$

The final estimates for the spatial locations, $\hat{s}_V$ and $\hat{s}_A$, will depend on the posterior probability for a common or separate events, $p(C \mid x_V, x_A)$. If the objective is to minimize mean squared error of the spatial estimates,

$$(\hat{s}_V - s_V)^2 + (\hat{s}_A - s_A)^2,$$

then the optimal estimates are obtained by a model averaging strategy (Körding *et al.*, 2007):

$$\hat{s}_A = p(C = 1 \mid x_V, x_A)\hat{s}_{A,C=1} + [1 - p(C = 1 \mid x_V, x_A)]\hat{s}_{A,C=2} \tag{14}$$

$$\hat{s}_V = p(C = 1 \mid x_V, x_A)\hat{s}_{V,C=1} + [1 - p(C = 1 \mid x_V, x_A)]\hat{s}_{V,C=2}. \tag{15}$$

Note that these overall estimates are no longer linear combinations of x_V and x_A as the posterior probabilities for common/separate events are not linear in these two variables. Marginalization yields the predicted distribution of visual positions that can be compared with experimental data:

$$p(\hat{s}_V \mid s_V, s_A) = \int\int p(\hat{s}_V \mid x_V, x_A)p(x_V \mid s_V)p(x_A \mid s_A)\mathrm{d}x_V\mathrm{d}x_A, \tag{16}$$

with an analogous expression for the auditory positions.

Two alternative decision strategies have been investigated in Wozny *et al.* (2010). In the model selection strategy, the location estimate on each trial is based purely on the causal structure that is more likely, given the sensory evidence and prior bias about the two causal structures; for example, for the auditory estimate,

$$\hat{s}_A = \begin{cases} \hat{s}_{A,C=1} & \text{if } p(C = 1 \mid x_V, x_A) > .5; \\ \hat{s}_{A,C=2} & \text{if } p(C = 1 \mid x_V, x_A) \leq .5. \end{cases} \tag{17}$$

This coincides with the Bayes optimal strategy with two classes.

In the probability matching strategy, the selection criterion is stochastic based on matching the posterior probability of the structure by sampling from a uniform distribution on each trial:

$$\hat{s}_A = \begin{cases} \hat{s}_{A,C=1} & \text{if } p(C=1 \mid x_V, x_A) > u \\ & \text{where } u \text{ is from a standard uniform distribution;} \\ \hat{s}_{A,C=2} & \text{if } p(C=1 \mid x_V, x_A) \leq u. \end{cases} \tag{18}$$

When it comes to finding out which of these three strategies are actually realized in a given context, experimental data are often somewhat ambiguous. In a large-scale study by Wozny and collaborators (Wozny *et al.*, 2010), a majority of subjects was best described by the probability matching strategy. This may seem surprising as it is clearly suboptimal compared with the model selection strategy. The authors argue, however, that observers typically try to find regularities in random patterns, so that the probability matching strategy is only suboptimal in a static environment and may be closer to optimality when predictable patterns actually exist (Vulkan, 2000).

On the other hand, further evidence for model averaging comes from Rohe & Noppeney (2015b) using four different levels of visual reliability. Participants responded to two questions presented sequentially: First, they localized the auditory spatial signal as accurately as possible by pushing one of five buttons that corresponded spatially to the stimulus locations. Second, they decided whether the visual and auditory signals were generated by common or independent sources and indicated their response via a two-choice key press. Rohe and Noppeney concluded that participants obtained an auditory localization estimate by combining the spatial estimates under the two causal structures weighted by their posterior probabilities, thus favoring the model averaging strategy. Moreover, in the judgment task participants reported a common source if the common-source posterior probability was larger than 0.5. Thus, in neither task did participants employ suboptimal sampling strategies.

As pointed out by one of the reviewers, the role of causal inference in modality combination other than visual–auditory is less clear. For example, de Winkel *et al.* (2015) observed subjects integrating visual and inertial information in heading estimation even under large spatial discrepancies (forced fusion). In a later study (de Winkel *et al.*, 2017), they did find evidence for causal inference models, which, however, also depended in a complex way on the specifics of the motion profiles (acceleration, maximum velocity, duration) of the stimuli.

Combining Bayesian causal inference with fMRI

In a landmark study by Rohe & Noppeney (2015a), participants localized audiovisual signals that varied in spatial discrepancy and visual reliability, while their brain activity was measured using functional magnetic resonance imaging (fMRI). As described in Kayser & Shams (2015), they first fit the causal inference model to the perceptual data to investigate the mapping between brain activity and the different spatial estimates predicted by the model. The estimates were predicted by either unisensory input (corresponding to the distinct causal origins hypothesis), by the fusion of the two sensations (corresponding to the single causal origin hypothesis), or by the causal inference model (the weighted combination of fusion and segregation). For that they had to link the selectivity to spatial information reflected in distributed patterns of fMRI activity to the spatial estimates predicted by each model component.

The authors concluded from their analyses that Bayesian causal inference is performed by a hierarchy of multisensory processes. At the bottom of the hierarchy, in auditory and visual areas, location is represented on the basis that the two signals are generated by independent sources. At the next stage, in posterior intraparietal sulcus, location is estimated under the assumption that the two signals are coming from a common source. Only at the top of the hierarchy, in anterior intraparietal sulcus, the uncertainty about the causal structure of the world is taken into account and sensory signals are combined as predicted by Bayesian causal inference.

Bayesian coupling prior model and remapping

In contrast to the standard cue combination model of the first section, the causal inference models of the last section do not have to assume mandatory fusion in the presence of large discrepancies between the perceptual estimates. By assuming unbiasedness, however, these models still cannot determine how much of a large discrepancy is due to the presence of noise or due to the difference in bias. The next model class to be described, the Bayesian coupling prior model suggested by Ernst and colleagues (see Ernst, 2012, for an overview and discussion), addresses this issue by introducing a bivariate prior probability distribution representing the co-occurrence statistics for the estimates.

In this section, we consider the case of visual and haptic estimates adopted from the Ernst notation and also follow closely the description in Ernst (2012). Bias is assumed to be additive; that is, the biased sensory signals are

$$S_V = S_{W_V} + B_V \quad \text{and} \quad S_H = S_{W_H} + B_H, \tag{19}$$

with S_{W_V} and S_{W_H} denoting the world attributes (like size) from which the sensory signals S_V and S_H are derived. The biases B_V and B_H may, for example, result from optical distortion or muscle fatigue when touching an object. It is first assumed that a single world attribute is sensed by vision and touch such that $S_{W_V} = S_{W_H}$. Both S_V and S_H are corrupted by independent Gaussian noise (ϵ) with variance σ_V^2 and σ_H^2, respectively. So

$$z_V = S_V + \epsilon_V \quad \text{and} \quad z_H = S_H + \epsilon_H.$$

The joint likelihood function of (z_V, z_H) is a bivariate normal $\mathcal{N}((S_V, S_H), \Sigma_z)$,

$$p(z_V, z_H \mid S_V, S_H) = \frac{1}{N_1} \exp\left[-\frac{1}{2}\left(\frac{(z_V - S_V)^2}{\sigma_V^2} + \frac{(z_H - S_H)^2}{\sigma_H^2}\right)\right] \tag{20}$$

with a normalization factor N_1 and covariance matrix

$$\Sigma_z = \begin{pmatrix} \sigma_V^2 & 0 \\ 0 & \sigma_H^2 \end{pmatrix}.$$

The goal is to get an estimate of the biases B_V and B_H in order to estimate the world attribute $S_{W_V} = S_{W_H}$ by subtraction in Eqn (19). The index MLE (maximum-likelihood estimate) refers to the estimates as inferred directly from the sensory signals. The discrepancy between the signal estimates,

$$\hat{D}^{\text{MLE}} = \hat{S}_V^{\text{MLE}} - \hat{S}_H^{\text{MLE}},$$

is the result of both the difference in the biases of the signals and the difference as a result of the noise in the sensory estimates. As

argued in Ernst (2012), the Bayesian approach resolves this ambiguity by assuming that the perceptual system has acquired a priori knowledge about the co-occurrence of the sensory signals S_V and S_H reflected by a prior distribution $p(S_V, S_H)$. This coupling prior can be interpreted as a predictor of how strongly one sensory signal reflects what the other one should be. The coupling prior model takes the product of the prior and the likelihood to compute the a posteriori distribution using Bayes's rule:

$$p(S_V, S_H) \times p(z_V, z_H \mid S_V, S_H) \propto p(S_V, S_H \mid z_V, z_H).$$

The prior distribution is assumed to be a bivariate normal with mean (0, 0). First, we consider the special case of unbiasedness, in addition to $S_{W_V} = S_{W_H}$. Thus, the distribution of joint signals would concentrate on the first diagonal (top row of Fig. 1), a special case of the upper Fréchet bound (Mari & Kotz, 2001, p. 155). This singular distribution can be approximated by

$$p(S_V, S_H) = \frac{1}{N_2} \exp\left\{-\frac{1}{2}\frac{(S_V - S_H)^2}{\sigma_x^2}\right\} \qquad (21)$$

with $\sigma_x \to 0$. Here, σ_x^2 measures the spread of the joint distribution function (middle row of Fig. 1) and N_2 is another normalization constant. According to the Bayes's rule, the posterior distribution will also have probability mass only on the diagonal. Because of unbiasedness, the maximum a posteriori (MAP) estimate for the posterior distribution is shifted, relative to the MLE of the likelihood function, in the direction of α (arrow in Fig. 1, top right) determined by the weights of Proposition 1 of the standard cue combination model,

$$\alpha = \arctan\left(\frac{w_V}{w_H}\right) = \arctan\left(\frac{\sigma_H^2}{\sigma_V^2}\right).$$

Next, we consider the case with a potential additive bias as in Eqn (19) and $S_{W_V} = S_{W_H}$. The prior is the same as in Eqn (21) where $\sigma_x^2 \neq 0$ corresponds to the variance of the joint distribution of biases. Again using Bayes's rule to combine the joint likelihood function with the prior gives rise to a MAP estimate of the sensory signals, $(\hat{S}_V^{\mathrm{MAP}}, \hat{S}_H^{\mathrm{MAP}})$ as well as the best possible estimate of the discrepancy as a result of the bias $\hat{D}^{\mathrm{MAP}} = \hat{S}_V^{\mathrm{MAP}} - \hat{S}_H^{\mathrm{MAP}} = \hat{B}_V - \hat{B}_H$. The MAP estimate is shifted with respect to the likelihood $\hat{S}^{\mathrm{MLE}}$ (see arrow in Fig. 1, right column). Its direction α is again determined by the variances σ_V^2 and σ_H^2. The length of the arrow indicates the strength of the integration; it is determined by a weighting of the likelihood function and the prior (conditioned on the line through the MLE and MAP

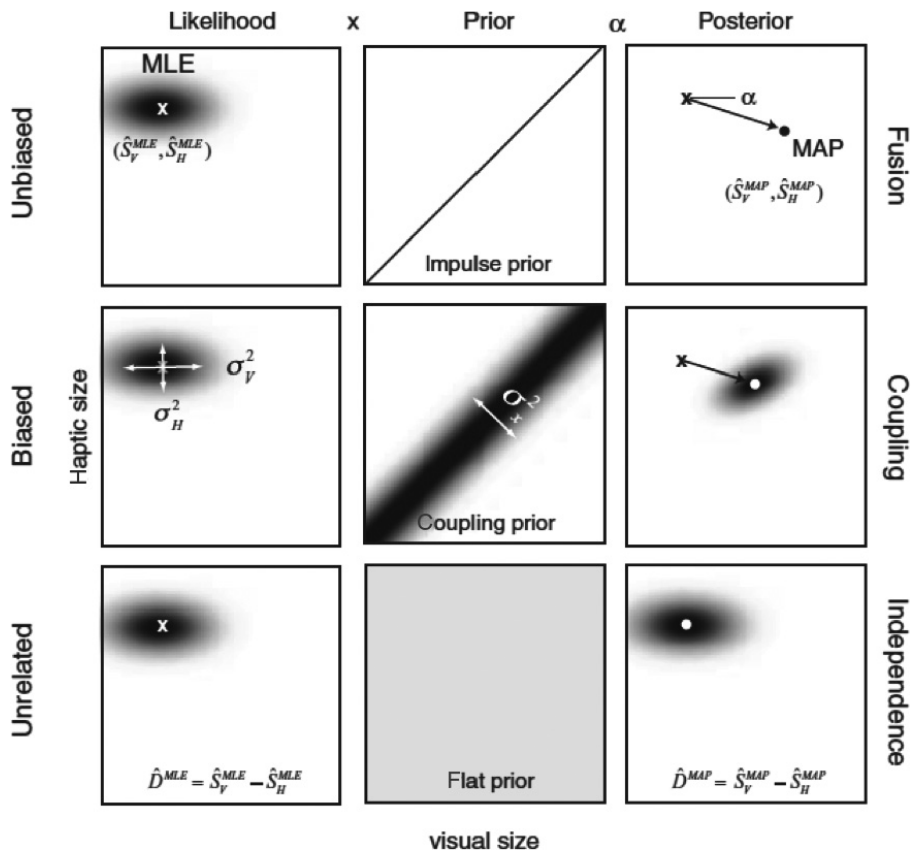

FIG. 1. The combination of visual and haptic measurements with different prior distributions. (Left column) Likelihood functions resulting from noise with standard deviation σ_V twice as large as σ_H; × marks the maximum-likelihood estimate (MLE) of the sensory signals $(\hat{S}_V^{\mathrm{MLE}}, \hat{S}_H^{\mathrm{MLE}})$. (Middle column) Prior distributions aligned with the positive diagonal (i.e., $S_{W_V} = S_{W_H}$) with different variances σ_x^2. (Top) Impulse prior $\sigma_x^2 = 0$. (Middle) Intermediate coupling prior $0 < \sigma_x^2 < \infty$. (Bottom) Flat prior ($\sigma_x^2 \to \infty$). (Right column) Posterior distributions, which are the product of the likelihood and prior distributions according to Bayes's rule. The dot marks the maximum a posteriori (MAP) estimate $(\hat{S}_V^{\mathrm{MAP}}, \hat{S}_H^{\mathrm{MAP}})$. Black arrows correspond to the shift in the MAP estimate relative to the MLE. The orientation of the arrows α correspond to the weighting of the $\hat{H}_V$ and $\hat{H}_H$ estimates. The length of the arrow indicates the degree of coupling from fusion to independence (From: Stein, Barry E., ed., The New Handbook of Multisensory Processing, Figure 30.3, p. 531, © 2012 Massachusetts Institute of Technology, by permission of The MIT Press).

estimates, i.e., in direction α). Note that the higher the value of σ_x^2, the lower the weight of the prior because the discrepancy between the signals is caused more likely by a bias and not noise in the estimates.

Finally, if the subject has no prior knowledge about the joint distribution of the stimuli in the world, the prior is flat, that is, thus noninformative, and the MAP estimate is given by the maximum-likelihood estimate (MLE) of distribution (20) (bottom row of Fig. 1),

$$\left(\hat{S}_V^{\mathrm{MAP}}, \hat{S}_H^{\mathrm{MAP}}\right) = \left(\hat{S}_V^{\mathrm{MLE}}, \hat{S}_H^{\mathrm{MLE}}\right)$$

which corresponds to its mean.

The author goes on to show that combining a (Gaussian) bivariate prior probability distribution of bias, $p(B_V, B_H)$, centered at (0, 0) and with variances independent of (S_V, S_H), with the likelihood function according to Bayes's rule, will lead to a bias estimate $(\hat{B}_V^{\mathrm{MAP}}, \hat{B}_H^{\mathrm{MAP}})$ (for details we refer to Ernst, 2012; Ernst & Di Luca, 2011).

The predictions of this model, regarding both bias and variance, have been confirmed, for example, by an experimental study of perceived numerosity of visual and haptic events (Bresciani *et al.*, 2006). Modifications of the model, relaxing some of the assumptions, for example, allowing for correlated noise distributions of the perceptual estimates are presented in Ernst (2012) as well. A recent study in tool use found clear evidence of reliability-based weighting while the variability in the biased position judgments was systematically larger than the optimal variability as predicted by the sensory coupling model (Debats *et al.*, 2017).

Some criticism of cue combination models

Rosas & Wichmann (2011) raise two important points of criticism against cue combination models, both based on statistical arguments. First, identification of cue reliability with variance neglects the fact that (sample) variance, and for that matter, all moment-based statistics are highly sensitive to outliers: "Thus calling the minimum-variance unbiased estimator "optimal" is an unfortunate choice – optimality implies a teleological correctness that the minimum-variance unbiased estimator cannot deliver." (Rosas & Wichmann, 2011, p.147). This becomes evident in the case of heavy-tailed distributions like the Cauchy that do not have a finite (population) variance. Second, while the authors acknowledge that cue combination may be sensitive to the reliability of the cues, inferring "optimality" in the minimum-variance-unbiased-estimator sense is seen as problematic for statistical testing reasons: It requires affirmation of the null hypothesis, and there may only be a small range of an appropriate sample size to yield enough power for rejecting it and, simultaneously, avoiding the accusation that rejection is trivial because of too large a sample size. Thus, the often claimed "consistency" with statistically optimal behavior is weak evidence, and the authors recommend, for example, to study under what circumstances behavior may actually result from simpler heuristic "nonoptimal" cue combination rule.

Crossmodal recalibration

The notion of cue reliability has been identified as a fundamental factor in cue combination models. In addition to reliability (precision), a veridical representation of a stimulus complex is expressed by its accurate localization in space and time. Deviation from perfect accuracy is measured as bias of a percept, an instance of which had already seen above in the Bayesian coupling prior model. The models to be briefly discussed now are not only concerned with immediate effects of presenting crossmodal stimuli but try to predict aftereffects, that is, whether the modalities are readjusted or recalibrated relative to each other for some time. Studies on recalibration abound, with emphasis on visual–auditory interactions, but modeling efforts are somewhat scarce (see Chen & Vroomen, 2013, for a comprehensive review of empirical and theoretical findings).

For stimuli presented at different levels of spatial discrepancy (in particular, ventriloquism), modeling relies once again on Bayesian inference and optimal cue combination (Alais & Burr, 2004; Körding *et al.*, 2007; Shams & Beierholm, 2010). Except for Ernst & Di Luca (2011) (see also Ernst, 2012), none of these explicitly develops the dynamics of aftereffects as function of time, that is, from time point t to $t + 1$.

Somewhat less well-known is modeling the phenomenon of temporal ventriloquism where temporal aspects of a visual stimulus, that is, onset, interval, or duration, can be shifted by slightly asynchronous auditory stimuli (see Chen & Vroomen, 2013). Two different types of temporal ventriloquism have been observed. Lag adaptation refers to the observation that subjects perceive audiovisual stimuli as simultaneous when they are repeatedly presented, separated a by fixed temporal interval (Fujisaki *et al.*, 2004). This still holds when the stimulation interval is not fixed but sampled from a Gaussian distribution with a positive peak such that light precedes sound by 80 ms on average. On the other hand, lag adaptation was not observed in judging the order of two tactile stimuli, each of which delivered to either the left or right hand. Rather, participants showed completely opposite perceptual changes that conformed to a Bayesian integration theory in that two simultaneous stimuli were perceived as having occurred in the order of the most frequent lag (Miyazaki *et al.*, 2006). The authors referred to this type of temporal ventriloquism as Bayesian calibration.

In their Bayesian integration model, Miyazaki *et al.* (2006) show how the final shift in the point of simultaneity can be expressed as the sum of two counteracting terms, one representing the Bayesian calibration and the other reflecting the effect of lag adaptation, irrespective of whether the lag adaptation occurs before or after Bayesian calibration. The model assumes that subjects base their judgments on the estimation of the stimulation interval ($t_{\mathrm{estimated}}$), which is estimated from the sensed stimulation interval (t_{sensed}) that is assumed to follow a Gaussian distribution around the true stimulation interval t_{true} with a standard deviation σ_{sensed}. By Bayesian rule,

$$p(t_{\mathrm{true}} \mid t_{\mathrm{sensed}}) = p_{\mathrm{prior}}(t_{\mathrm{true}})p(t_{\mathrm{sensed}} \mid t_{\mathrm{true}})/p(t_{\mathrm{sensed}}),$$

where $p_{\mathrm{prior}}(t_{\mathrm{true}})$ denotes the prior probability of the true interval, distributed as Gaussian with mean and variance $(\mu_{\mathrm{prior}}, \sigma_{\mathrm{prior}}^2)$. Adding further assumptions (see supplement of Miyazaki *et al.*, 2006), the point of simultaneity d_u can be calculated as

$$d_u = -\left(\frac{\sigma_{\mathrm{sensed}}}{\sigma_{\mathrm{prior}}}\right)^2 \mu_{\mathrm{prior}} + \frac{\sigma_{\mathrm{prior}}^2 + \sigma_{\mathrm{sensed}}^2}{\sigma_{\mathrm{prior}}^2} c.$$

When the size of lag adaptation, c, moves from zero to μ_{prior}, the point of simultaneity moves from full Bayesian calibration (first term) to full lag adaptation (μ_{prior}). A later paper from the same laboratory (Yamamoto *et al.*, 2012), showing that lag adaptation in audiovisual temporal order judgment is pitch-insensitive whereas Bayesian calibration is pitch-specific, supported the hypothesis that Bayesian calibration operates for audiovisual stimuli as well but is concealed behind the lag adaptation mechanism.

In a similar, computational Bayesian approach combining both types of calibration, Sato and Aihara extended their previous model (Sato *et al.*, 2007) assuming that lag adaptation represents learning of the likelihood function and Bayesian calibration represents learning of the prior distribution (Sato & Aihara, 2011). They present an explicit updating rule when an observer estimates the mean value of the likelihood function and of the prior distribution.

It should be noted here that most studies of temporal ventriloquism define the point of simultaneity as the one where the psychometric function (of temporal order) intersects the probability at 0.5. However, when at times a 3-alternative choice is offered, including judgment of "simultaneity", results may not be consistent with those from a 2-alternative choice paradigm. This problem is addressed in a common framework developed in García-Pérez & Alcalá-Quintana (2012).

Finally, from a more general point of view, Zaidel *et al.* (2013) stress the distinction between "supervised" calibration, where environmental feedback regarding cue accuracy is given, and "unsupervised" calibration, where such feedback is absent. Using discrepant visual and vestibular motion stimuli, they measured the combined influence of cue accuracy and cue reliability. When the less reliable cue was inaccurate, it alone got calibrated. However, when the more reliable cue was inaccurate, cues were combined and calibrated together in the same direction. A computational model similar to the one proposed in Ernst & Di Luca (2011) for unsupervised calibration, in which supervised and unsupervised calibration work in parallel, where the former only relies on the multisensory percept, but the latter can calibrate cues individually, accounted for their data best.

Correlation detector models

Although this model class could formally be subsumed under cue combination, its members have a number of features that warrant a separate discussion. At the level of basic correlation detector units, these models address the correspondence problem—that is, whether information from different modalities originate from the same event and should be integrated—by computing the correlation and temporal lag between streams of stimulus information from two different sensory modalities. The rationale is that, when signals from different modalities are actually coming from the same event, they cross-correlate in both time and space. It has indeed been shown that multisensory cue integration is statistically optimal (in the sense discussed above) only when signals are temporally correlated (Parise *et al.*, 2012, 2013). The principle of correlation detection is well established in several areas of neuroscience, including vision (Ohzawa, 1998) and hearing (Jeffress, 1948).

The multisensory correlation detector (MCD) model

An early model of multisensory synchrony detection relying on cross-correlation operations was proposed in Fujisaki & Nishida (2007). Up to date, the most comprehensive correlation detector model for multisensory integration has been developed by Parise and colleagues (Parise & Ernst, 2016).. It is based on the Hassenstein–Reichardt model for local motion detection (Hassenstein & Reichardt, 1956), commonly referred to as the Reichardt detector (van Santen & Sperling, 1985).

The Reichardt detector consists of two mirror-symmetrical subunits which have a receptor that can be stimulated by an input (light in the case of visual system). In each subunit, when an input is received, a signal is sent to the other subunit. At the same time, the signal is delayed within the subunit and, after this temporal filter, is then multiplied by the signal received from the other subunit. The multiplied values from the two subunits are then subtracted to produce an output (see Fig. 2). The preferred direction is determined by whether the difference is positive or negative. The direction which produces a positive outcome is the preferred direction. Various elaborations of the basic Reichardt model have been proposed to accommodate this motion detection scheme to perform in a species-specific way, for example Fisher *et al.* (2015), and Borst & Euler (2011) for a review.

Instead of receiving unisensory (visual) information from neighboring receptive fields, the multisensory correlation detector (MCD) model by Parise and Ernst receives inputs from spatially aligned visual and auditory receptive fields. In a first step, the time-varying visual and auditory signals $S_V(t)$ and $S_A(t)$ undergo separate temporal low-pass filters that account for the specific response characteristics of each sense (transduction, transmission and early unisensory processing). The filters are the same as in Burr *et al.* (2009), that is,

$$f_{\mathrm{mod}}(t) = t\exp(-t/\tau_{\mathrm{mod}}),$$

where mod stands for the visual, auditory, or visual–auditory modality and τ_V, τ_A are the respective temporal constants of the filter. The next stage parallels the Reichardt detector in that these filtered signals are fed into two mirror-symmetrical subunits, which multiply the signals after introducing another temporal shift to one of them through another low-pass filter of the same type as before with constant τ_{VA}, identical for both subunits (see Fig. 2a). Specifically, for subunits (u_1, u_2) and signals ($S_V(t)$, $S_A(t)$)

$$u_1(t) = \{[S_A(t) * f_A(t)] * f_{VA}(t)\} \times [S_V(t) * f_V(t)]$$
$$u_2(t) = [S_A(t) * f_A(t)] \times \{[S_V(t) * f_V(t)] * f_{VA}(t)\}$$

where * stands for convolution with the low-pass temporal filters. In the next step, the responses of the subunits are multiplied or subtracted to obtain:

$$\mathrm{MCD}_{\mathrm{Corr}}(t) = u_1(t) \times u_2(t)$$
$$\mathrm{MCD}_{\mathrm{Lag}}(t) = -u_1(t) + u_2(t),$$

which are time-varying responses representing the local temporal correlation and lag across the signals. These responses are turned into a single summary variable representing the amount of evidence from each trial by simply averaging the output of the detectors over a long-enough time window.

Here, correlation is calculated by multiplying the outputs of the two subunits, and temporal lag is detected by subtracting the outputs of the subunits, like in the Reichardt detector. This yields an output with a sign that represents the temporal order of the signals (see Fig. 2b).

Parise & Ernst (2016) tested the MCD model with a forced-choice dual task: Stimuli consisted of sequences of five visual and auditory impulses presented over an interval of 1 s, and subjects had to report (i) whether or not the visual and auditory sequences appeared to be causally related and formed a perceptual unity and (ii) the relative temporal order of the two sequences. Notably, the temporal structures of the visual and auditory signals were generated independently and varied randomly across trials; thus, there was no unambiguous information to decide about the relative temporal order or about the common causal structure of the signals. Subjects' data were analyzed using psychophysical reverse correlation techniques. These analyses were performed on the cross-correlation profile of

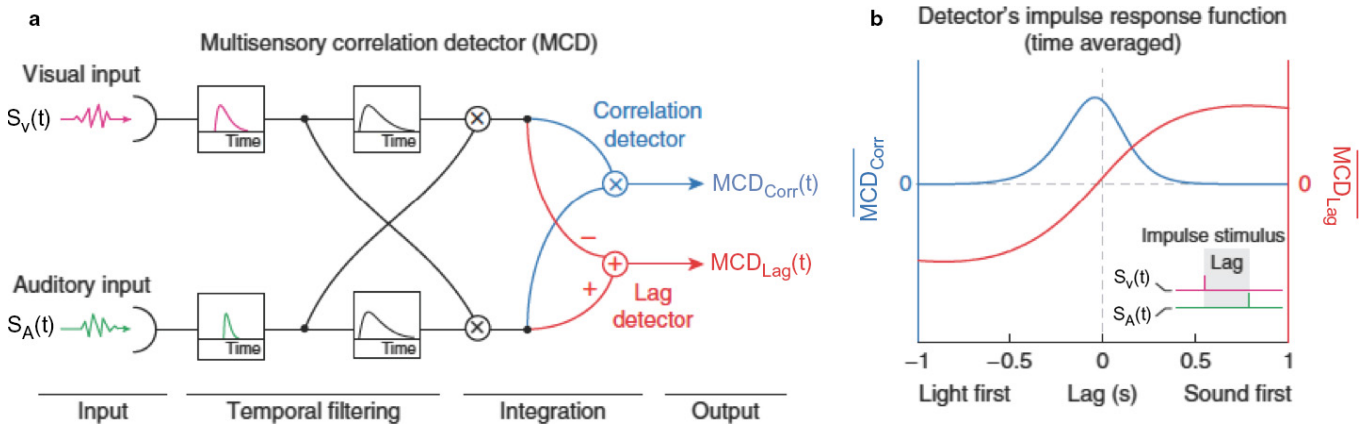

FIG. 2. Multisensory correlation detector (MCD) model. (a) Schematic representation of the model. (b) Time-averaged impulse response function of the MCD (see text). (Reprinted with permission from Nature Communications, Parise & Ernst, 2016).

the signals as it provides a measure of similarity of the signals as a function of lag, and it highlights common features across complex signals. In fact, signals with high correlation at short lags were more likely perceived as sharing a common cause, while responses in the temporal order judgments were driven by the sign of the lag at maximum cross-correlation (i.e., light vs. sound lead). To compare the subject data with predictions of the MCD model, the same signals used in the experiment were fed into the model to perform reverse-correlation analyses on the model responses as well. The authors could show that the model could near perfectly reproduce the shapes of both empirical classification images (for details, see Parise & Ernst, 2016). Keeping the same temporal constants, the MCD model could also fit data from previous temporal order judgment, synchrony detection and judgment, and audiovisual correspondence detection studies.

Despite the originality and impressive fit to data of the MCD model, we also see some issues and possible shortcomings of the approach. This study, as well as previous ones like Parise *et al.* (2012), makes a clear case that, in a noisy environment, the brain tends to infer causality (e.g., unity of events) from correlation. Detecting correlation between stimuli from different sensory modalities takes time, and this is reflected by the long time constant for the estimate of audiovisual filter lag ($\tau_{VA} = 786$ ms); on the other hand, there are many studies, including reaction time measurements, documenting that lawful multisensory integration processes occur at much shorter time scales, that is, between 100 and 400 ms. Second, the temporal correlation detector itself is mute with respect to the spatial configuration of the stimuli, whereas spatial arrangement between stimuli from different senses is known to be a critical factor in multisensory integration. Finally, due to MCD being rooted in linear systems theory, possible effects of nonlinear dependency between the senses, as arguably observable in motion stimuli, cannot be taken into account in the present version of the theory.

Models for reaction time paradigms

In the redundant-signals paradigm (RSP), stimuli from two (or more) different modalities are presented simultaneously, and participants are instructed to respond to a stimulus of any modality, whichever is detected first, typically by a button press. In the focused attention paradigm (FAP), one modality is declared the target modality to which the subject is asked to react as quickly as possible, by button press or an eye movement (saccade). Reaction time is measured from the onset of the first stimulus (for RSP), or the target modality stimulus (for FAP), to the response.

For each stimulus or crossmodal stimulus combination, we observe samples from a random variable representing the reaction time measured in a given trial. Let $F_V(t)$, $F_A(t)$, and $F_{VA}(t)$ denote the (theoretical) distribution functions of reaction time in a unimodal visual, auditory, or a visual–auditory context, respectively, when a specific stimulus or stimulus combination is presented. Note that, for simplicity, we write $F_V(t)$, etc., instead of $F_{\mathcal{V}}(t)$. Typically, mean time to respond in the crossmodal condition is shorter than that in either of the unimodal conditions in both RSP and FAP, but it also critically depends on the stimulus-onset time (SOA) between the signals, denoted as τ (e.g., Diederich & Colonius, 2004a).

Race model for reaction times

Let us assume that, in the crossmodal condition, (i) each individual stimulus elicits a process performed in parallel to the others and (ii) the finishing time of the faster process determines the observed RT. This is known as the "race model" for RTs (Raab, 1962). For simplicity, we here equate observable RTs with the finishing times of the "race", neglecting any addition components of RT like motor processing, etc.. Assuming random variability of the finishing times, mean RT in the crossmodal condition with SOA = τ, denoted as $RT_{VA+\tau}$, is predicted to be shorter than the faster of the unimodal mean RTs because, on a portion of the crossmodal trials, slow responses of the "faster" modality will be replaced by faster responses of the "slower" modality:

$$\mathbf{E}\mathrm{RT}_{VA+\tau} \equiv \mathbf{E}\min\{\mathrm{RT}_V, RT_{A+\tau}\} \leq \min\{\mathbf{E}\mathrm{RT}_V, \mathbf{E}\mathrm{RT}_{A+\tau}\}. \quad (22)$$

The equivalence on the left-hand side of Eqn (22) expresses assumption (ii) of the race model, and the inequality follows as a special case of Jensen's inequality (e.g., Bhattacharya & Waymire, 1990). This is an effect of probability summation, or "statistical facilitation", and no "true" multisensory integration of the unisensory processes is required.

The context invariance assumption

Assumption (ii) of the race model involves another, often not explicitly recognized assumption of the model (but see Luce, 1986, pp. 129–130): Note that each context $\mathcal{V}, \mathcal{A}$, or $\mathcal{VA}$ refers to a different

sample space (and corresponding σ-algebra), as there is no experimental condition in which simultaneous realizations of the random variables RT_V, RT_A, and $RT_{VA+\tau}$ are observed. Thus, without further assumptions nothing can be said about their stochastic dependency. A possible solution is to claim that RT_V, RT_A, and $RT_{VA+\tau}$ be stochastically independent. In that case, the bivariate distribution of (RT_V, RT_A) equals the product of the univariate distributions,

$$\begin{aligned} H_{VA}(s,t) &\equiv \Pr(\mathrm{RT}_V \le s, \mathrm{RT}_A \le t) \\ &= F_V(s)F_A(t), \end{aligned}$$

for all non-negative s, t. In particular,

$$H_{VA}(s,\infty) = F_V(s) \text{ and } H_{VA}(\infty,t) = F_A(t). \tag{23}$$

For the stochastically independent race model,

$$\begin{aligned} F_{VA}(t) &= \Pr(\min\{\mathrm{RT}_V, \mathrm{RT}_A\} \le t) \\ &= \Pr(\mathrm{RT}_V \le t) + \Pr(\mathrm{RT}_A \le t) - \Pr(\mathrm{RT}_V \le t)\ \Pr(\mathrm{RT}_A \le t). \end{aligned} \tag{24}$$

Observing crossmodal RT distributions exceeding the right-hand side of Eqn (24) at any time point t leads to a rejection of the independent race model. Note, however, that the equalities in (23) can also hold without assuming stochastic independence. This amounts to claiming that the two marginal distributions of $H_{VA}(s, t)$ are equal to the univariate distributions $F_V(s)$ and $F_A(t)$, respectively. This assumption, known as context invariance (Colonius, 1990; Townsend & Nozawa, 1995), is not empirically testable because $H_{VA}(s, t)$ is not observable in paradigm RSP. As recently demonstrated in Miller (2016), context invariance is an essential part of the race model concept and dropping it makes the race model unfalsifiable.

The race model inequality

Rewriting Eqn (24) without the independence assumption,

$$F_{VA}(t) = \Pr(\mathrm{RT}_V \le t) + \Pr(\mathrm{RT}_A \le t) - H_{VA}(t,t). \tag{25}$$

Because $H_{VA}(t, t)$ is always non-negative, dropping it leads to inequality

$$\begin{aligned} F_{VA}(t) &\le \Pr(\mathrm{RT}_V \le t) + \Pr(\mathrm{RT}_A \le t) \\ &= F_V(t) + F_A(t). \end{aligned} \tag{26}$$

This is the race model inequality. As observed by Jeff Miller (Miller, 1982), it constitutes a stronger test of the model than a test based on the independent race model (Eqn 24) or on the level of means using inequality (22). It turns out that the upper bound $F_V(t) + F_A(t)$ corresponds to a race under maximal negative dependence between RT_V and RT_A (Colonius, 1990, 2016). Generalizations of the race model inequality to three different modalities have been discussed in (Diederich, 1992b; Colonius *et al.*, 2017).

The race model inequality has become the standard tool for probing whether observed reaction times to crossmodal stimuli are faster than predicted by a race mechanism. Gondan & Minakata (2016) discuss a variety of statistical methods in wide use. Many of these tests are somewhat questionable being based on multiple testing, and others are only applicable when data from several subjects are available. We have suggested an alternative, possibly less powerful, test based on comparing the area between $F_{VA}(t)$ and $F_V(t) + F_A(t)$ defined by all t values where the race model inequality is violated (Colonius & Diederich, 2006). It turns out this test simply compares $\mathbf{E}\mathrm{RT}_{VA}$ with the expected value of random variable min $\{V, A\}$ under maximal negative dependence, that is, SM^- in Eqn (4).

Violation of Eqn (26) at any time point t has been taken as evidence in favor of some form of multisensory integration taking place, above the level predicted by statistical facilitation. This conclusion may often be valid, in particular in a context with few simple, low-dimensional stimuli and under low noise levels. On the other hand, the race model test may sometimes be distorted by specifics of the experimental setup leading to serial dependency across trials. A prominent effect of trial history has been discussed as the "modality-switch" effect: Successive trials of the same modality may lead to faster and better performances (repetition priming) than when the modalities alternate in either uni- or bimodal trials (Gondan *et al.*, 2004). In a recent study employing a large set of natural images and sounds, Juan and colleagues found a large range of violations and nonviolations of the race model inequality as a function of the intrinsic physical parameters of the stimuli in both humans and monkeys (Juan *et al.*, 2017).

Coactivation models for reaction time

Coactivation models for reaction time have been defined in Miller (1982) as an alternative to race models, the latter being referred to as "separate activation models". Coactivation occurs when activation from different channels combines in "satisfying a single criterion for response initiation", whereas "in separate activation the system never combines activation from different channels in order to meet its criterion for responding" (ibid, p. 248). Coactivation models predict faster average reaction time to multiple stimuli compared to single stimuli because the combined activation reaches the criterion for response initiation faster.

The two most prominent classes of coactivation models differ in whether the accumulation of activation occurs in a discrete or continuous fashion.

Discrete accumulation: Counter models

It is assumed that presentation of a stimulus triggers a sequence of "events" occurring randomly over time. This is in analogy to a spiking neuron, but the model is formulated at a more abstract level. The only relevant property of the events is their time of occurrence, and all information about the stimulus is contained in the time course of the events. For example, the rate of the event sequence, that is, the mean number of events per unit time interval, is typically thought to be related to signal intensity.

Let $N_V(t)$, $N_A(t)$ denote the random number of events that have occurred by time t after visual or auditory stimulus presentation. Counter models assume that these numbers (counters) are internally represented over the course of time. Just as for the race model, an assumption of context invariance must be added when stimuli from different modalities are processed simultaneously. This amounts to assuming that the distribution of each of the counters $N_V(t)$, $N_A(t)$ is the same in uni- and bimodal conditions. Moreover, independence between the two is usually accepted as well.

If one were to postulate that a response is initiated as soon as any of the counters reaches a preset criterion, we would be back to the class of separate activation (race) models, with a race between counters taking place and the winner determining the response. In contrast, a coactivation model results if counters are combined, in particular,

$$N_{VA}(t) = N_V(t) + N_A(t).$$

Obviously, when two (or more) counters are added, a fixed criterion number of counts, c, will on average be reached faster than for any single counter. Usually, $N_V(t)$ and $N_A(t)$ are assumed to be stochastically independent for any t. To compute the distribution of the (random) waiting time S_c for the cth count to occur note the well-known identity (Bhattacharya & Waymire, 1990)

$$\Pr(S_c \leq t) = \Pr(N(t) \geq c).$$

The most tractable case for deriving exact quantitative predictions is the Poisson (counting) process where it is assumed that for each counter the times between successive events (interarrival times) are independent exponentially distributed random variables. Each Poisson process is characterized by a single constant, the intensity parameter λ. The expected waiting time for the cth count then simply is $\mathbf{E}S_c = c/\lambda$. Here, c is often interpreted as a bias parameter describing the subject's strategic behavior: When high accuracy is required, the subject may raise the criterion, minimizing erroneous responses, whereas, when high response speed is encouraged, c may take a lower value. It follows that the expected waiting time for the superposed Poisson process $N_{VA}(t)$ equals

$$\mathbf{E}S_c = c/(\lambda_V + \lambda_A).$$

Diederich (1992a, 1995) probed the Poisson counter model in the case of three modalities (visual, auditory, tactile). Specifically, rate parameters obtained in the trimodal condition were used to predict RTs in the bimodal conditions. Fits of the Poisson counter model to crossmodal RTs have generally been quite satisfying, at least at the level of RT means (Schwarz, 1989; Diederich & Colonius, 1991; Diederich, 1992a). If an additive (motor) component is added to the decision RT, variances can also be fit satisfactorily (Schwarz, 1994). Nevertheless, some features of this model are not consistent with observations. First, the model predicts that increasing stimulus intensity should lead to ever faster responses, without being able to account for any saturation effects. Second, it is not clear how the rule of "inverse effectiveness", according to which crossmodal facilitation is strongest when stimulus strengths are weak, can be generated by the superposition model. Finally, it seems that the counter model is, in principle, not able to predict inhibition of RT which may occur under certain crossmodal (spatial) configurations. These shortcomings of the superposition model have led to considering still another implementation of the coactivation idea.

Continuous accumulation: diffusion models

In the diffusion model for crossmodal RTs, counters are replaced by a two-dimensional Wiener process (or Brownian motion) $\{X_V(t), X_A(t), t \geq 0\}$ with drift parameters μ_V, μ_A, variance parameters σ_V^2, σ_A^2, correlation ρ, and initial conditions $X_V(t) = X_A(t) = 0$ (see, e.g., Schwarz, 1994). For nonoverlapping time intervals, the increments $\{X_i(t + h) - X_i(t), t, h \geq 0\}$ $(i = V, A)$ are independent, normally distributed, and independent of t. The decision to execute a response is made when the trajectory of the process in a given trial reaches an absorbing boundary c, which plays a role analogous to the criterion in the Poisson superposition model. The expected time to reach this threshold, the first passage time, is

$$\mathbf{E}T_i = c/\mu_i \text{ for } i = V \text{or} A,$$

independent of the variance. It can be shown that the superposed process, that is,

$$X_{VA}(t) = X_V(t) + X_A(t),$$

is again a Wiener process, with drift $\mu_V + \mu_A$ and variance $\sigma_V^2 + \sigma_A^2 + 2\rho\sigma_V\sigma_A$. Moreover, the expected time to reach boundary c equals

$$\mathbf{E}T_{VA} = c/(\mu_V + \mu_A).$$

Although the two models are fundamentally different from a theoretical viewpoint, predictions of the diffusion model and the Poisson superposition model are clearly indistinguishable as long as only expected decision (first passage) times are considered. Thus, the diffusion model was developed further in two different directions. Schwarz (1994) derived expressions for the first passage time as function of stimulus-onset asynchrony (SOA) and added a nonindependent motor component to obtain an excellent fit to the classic data set of Miller (1986). Using a matrix approximation approach to diffusion processes, Diederich (1992a, 1995) developed an Ornstein–Uhlenbeck process (OUP) model for crossmodal RTs. In OUP it is assumed that the drift rate, while constant over time, is a function of the activation level x:

$$\mu(x) = \delta - \gamma x,$$

where δ refers to the constant part of the drift driving the process to the criterion (absorbing boundary) and γ is a decay parameter. As before, δ is a monotonic function of stimulus intensity: Strong stimuli have large δ values implying that the trajectories first have a tendency to be steep and to quickly approach the boundary. However, for positive values of δ the drift $\mu(x)$ decreases the faster the larger the activation level x becomes, that is, the closer activation gets to the criterion. This is responsible for the trajectories to level off rather than to increase linearly over time. Moreover, when the stimulus signal is switched off, the drift becomes negative and activation is assumed to decay to its starting level, as δ takes on a value of zero. It is assumed that activation never drops below its initial level. This decay process, which cannot be represented in a superposition/counter model, has been discussed in studies of neuronal activity dynamics (Ricciardi, 1977; Tuckwell, 1995). As before in fitting the Poisson counter model, Diederich (1995) used parameters estimated from the trimodal condition to predict RTs in the bimodal conditions, with the trimodal drift rate defined as

$$\mu(x) = \delta_V + \delta_A + \delta_T - \gamma x.$$

By defining the drift rate in judicious ways, diffusion models afford some flexibility. For example, one way to generate RTs following the inverse-effectiveness rule is to define (Diederich & Colonius, 2004b)

$$\mu(x) = (\delta_V + \delta_A)[1 + (\delta_V^{\max} - \delta_V)(\delta_A^{\max} - \delta_A)],$$

with $\delta_V^{\max}, \delta_A^{\max}$ denoting drift rates at maximal (stimulus) intensity levels. This would yield an additive effect of intensity if at least one modality is at its maximum level, but an inverse-effectiveness effect if all stimuli are away from the their maximum levels.

Note that in the race and coactivation models discussed so far, error probabilities and error RTs have not been taken into account. This is not a problem as errors rates are typically kept very low in these RT experiments. On the other hand, it is plausible—not the least in real-world situations—that additional information about multisensory integration mechanisms and their optimality may be obtained when subjects are allowed to develop specific speed–accuracy trade-off strategies in responding to crossmodal stimuli. Both diffusion models and Poisson superposition models have been extended to predict error rates and error RTs in paradigms other than multisensory integration (e.g., Smith & van Zandt, 2000; van Zandt *et al.*, 2000; Diederich, 2003); thus, these models are quite promising in extending multisensory RT models in this direction. For example, a more elaborated version of drift was proposed in a recent diffusion model by Drugowitsch and colleagues for a visual–vestibular heading discrimination task predicting both correct and incorrect RTs (Drugowitsch *et al.*, 2014). To our knowledge, this is the first attempt to implement optimal cue integration, as a function of (time-varying) reliability of the unisensory inputs, into a crossmodal RT model. Subjects had to decide whether points of a random-dot 3D optic flow was moving rightward or leftward relative to straight ahead. With x_{vis} and x_{vest} denoting the visual and vestibular heading evidence, the drift rate is not simply equal to their sum. Optimal weighting of the visual and vestibular trajectories of particles moving between an upper and a lower bound results in

$$x_{\text{comb}} = \sqrt{\frac{k_{\text{vis}}^2(c)}{k_{\text{vis}}^2(c)+k_{\text{vest}}^2}}x_{\text{vis}} + \sqrt{\frac{k_{\text{vest}}^2}{k_{\text{vis}}^2(c)+k_{\text{vest}}^2}}x_{\text{vest}},$$

where $k_{\text{vis}}(c)$ indicates the sensitivity to the visual cue, depending on motion coherence c, and k_{vest} sensitivity to the vestibular cue (for details, we must refer to Drugowitsch *et al.*, 2014). As discussed by the authors, one strong assumption of the model is that subjects can somehow estimate their sensitivity to the cues in order to optimally weight them appropriately. They conclude from their model fitting that it quantitatively explained subjects' choices and reaction times, supporting the hypothesis that subjects accumulate evidence optimally over time and across sensory modalities, even when the reaction time is under the subject's control.

TWIN model

The final modeling approach to be discussed is the authors' time window of integration (TWIN) model. We consider it a framework within which a number of alternative quantitative models can be formulated rather than a specific parametrized model.

The TWIN model (Colonius & Diederich, 2004; Diederich & Colonius, 2004b) was developed to predict the effect of the spatiotemporal parameters of a crossmodal experiment on saccadic or manual RT. It postulates that a crossmodal (audiovisual) stimulus triggers a race mechanism in the very early, peripheral sensory pathways (first stage), followed by a compound stage of converging subprocesses comprising neural integration of the input and preparation of a response. This second stage is defined by default: It includes all subsequent, possibly temporally overlapping, processes that are not part of the peripheral processes in the first stage. The central assumption concerns the temporal configuration needed for crossmodal interaction to occur:

> TWIN assumption: *Crossmodal interaction occurs only if the peripheral processes of the first stage all terminate within a given temporal interval, the time window of integration.*

Thus, the window acts as a filter determining whether afferent information delivered from different sensory organs is registered close enough in time to trigger multisensory integration. Passing the filter is necessary, but not sufficient, for crossmodal interaction to occur because the amount of interaction may also depend on many other aspects of the stimulus context, in particular the spatial configuration of the stimuli. The amount of crossmodal interaction manifests itself in an increase, or decrease, of second-stage processing time. Although this amount does not directly depend on the stimulus-onset asynchrony (SOA) of the stimuli, temporal tuning of the interaction occurs because the probability of integration is modulated by the SOA value. The race in the first stage of the model is made explicit by assigning statistically independent, non-negative random variables V and A, say, to the peripheral processing times for a visual and an acoustic stimulus, respectively. With τ as SOA value and ω as (integration) window width parameter, the TWIN assumption implies that the event that multisensory integration occurs, denoted by I, equals

$$\begin{aligned} I &= \{|V-(A+\tau)|<\omega\} \\ &= \{A+\tau<V<A+\tau+\omega\} \cup \{V<A+\tau<V+\omega\}, \end{aligned}$$

where the presentation of the visual stimulus is arbitrarily defined as the physical zero time point. Thus, the probability of integration to occur, $P(I)$, is a function of both τ and ω, and it can be determined numerically once the distribution functions of A and V have been specified.

In empirical studies probing various aspects of TWIN (Diederich & Colonius, 2007a,b, 2008a,b), the peripheral processing times have been assumed to follow exponential distributions, mainly for ease of computation. However, adding a Gaussian component as second-stage processing time results in predicting RT distributions in the form of (a mixture of) ex-Gaussian distributions, which are quite common in RT modeling (Luce, 1986).

Writing S_1 and S_2 for first- and second-stage processing times, respectively, overall expected reaction time in the crossmodal condition with an SOA equal to τ, $\mathbf{E}[\text{RT}_{V\tau A}]$, is computed conditioning on event I (integration) occurring or not,

$$\begin{aligned} \mathbf{E}[\text{RT}_{V\tau A}] &= \mathbf{E}[S_1] + P(I)\mathbf{E}[S_2|I] + [1-P(I)]\mathbf{E}[S_2|I^c] \\ &= \mathbf{E}[S_1] + E[S_2|I^c] - P(I)\times\Delta. \\ &= \mathbf{E}[\min(V, A+\tau)] + \mu - P(I)\times\Delta. \end{aligned} \tag{27}$$

Here, I^c denotes the complementary event to I, μ is short for $\mathbf{E}[S_2|I]$, and Δ stands for $\mathbf{E}[S_2|\text{t } I^c] - \mathbf{E}[S_2|I]$. The term $P(I)\times\Delta$ is a measure of the expected amount of crossmodal interaction in the second stage, with positive Δ values corresponding to facilitation, negative ones to inhibition. Explicit expression for $\mathbf{E}[\text{RT}_{VA,\tau}]$ in the exponential version as a function of the parameters can be found in the cited references.

Obviously, event I cannot occur in the unimodal (visual or auditory) condition; thus, expected reaction time for these conditions is, respectively,

$$\mathbf{E}[\text{RT}_V] = \mathbf{E}[V] + \mathbf{E}[S_2|I^c] \text{ and } \mathbf{E}[\text{RT}_A] = \mathbf{E}[A] + \mathbf{E}[S_2|I^c].$$

Note that the race in first stage produces a not directly observable statistical facilitation effect (SFE) analogous to the one in the "classic" race model (Raab, 1962):

$$\text{SFE} \equiv \min\{\mathbf{E}[V], \mathbf{E}[A]+\tau\} - \mathbf{E}[\min\{V, A+\tau\}].$$

This contributes to the overall crossmodal interaction effect predicted by TWIN, which amounts to:

$$\min\{\mathbf{E}[\mathrm{RT}_V], \mathbf{E}[\mathrm{RT}_A] + \tau\} - \mathbf{E}[\mathrm{RT}_{VA,\tau}] = \mathrm{SFE} + P(I) \times \Delta.$$

Thus, crossmodal facilitation observed in a redundant-signals task may be due to multisensory integration or statistical facilitation, or to both. Moreover, a potential multisensory inhibitory effect occurring in the second stage may be weakened, or even masked completely, by simultaneous presence of statistical facilitation in the first stage. This shows that the TWIN framework combines features of the race model class with those of the class of coactivation models.

"Temporal window of integration" has become an important concept in describing crossmodal binding effects as function of age, specific disorders, and training in a variety of multisensory integration tasks apart from RTs, for example, temporal order judgment (TOJ) or simultaneity (synchronicity) judgment (Vroomen & Keetels, 2010; for a recent review). As recently shown (Diederich & Colonius, 2015), a simple extension of the TWIN approach to RTs affords a common theoretical basis for this concept across these different paradigms by tying it to a common model. In this extension, window width emerges as a model parameter controlling, on the one hand, the probability of crossmodal interaction occurring in reaction time and, on the other, the probability of judging the temporal order of the stimuli. In the TOJ task, the width of the window determines how often the two stimuli will be "bound together" and, thereby, how often the subject can only guess that the visual stimulus occurred first, requiring the introduction of a response bias parameter into the model. Widening the temporal window of integration in a RT task, or narrowing it in a TOJ task, can be seen as an observer's strategy to optimize performance in an environment where the temporal structure of sensory information from separate modalities provides a critical cue for inferring the occurrence of crossmodal events. Our re-analysis (Diederich & Colonius, 2015) of data from Mégevand *et al.*, (2013) supported their hypothesis of a smaller time window for the TOJ task compared to the crossmodal RT task. Recently, effects of week-long synchronous and asynchronous adaptation conditions on reaction times to audiovisual stimuli have been found (Harrar *et al.*, 2016). On the other hand, rapid recalibration taking place from one trial to the next would clearly be advantageous in a dynamically changing environment. This has actually been observed within an auditory localization task, where spatial recalibration occurred as a function of audiovisual discrepancy after a single trial presentation (Wozny & Shams, 2011; Mendonca *et al.*, 2015). The TWIN modeling framework may have to be extended in order to capture dependencies across trials. Clearly, further research on the mechanisms underlying both fast and slow recalibration is called for that will integrate RT adaptation effects with the (Bayesian) modeling approaches reviewed in the above section on crossmodal recalibration.

Computational neuronal models

Models in this class refer explicitly to multisensory mechanisms performed by single neurons or neural populations. This class is rapidly growing, and we will focus here only on a few approaches that try to relate them to behavioral performance, in particular cue integration (for a more comprehensive recent review, see Fetsch *et al.*, 2013). Most models so far are either referring to neurons in the deep layers of superior colliculus (dSC) or the dorsal medial superior temporal area (MSTd).

Probabilistic population codes

The probabilistic population codes (PPC) framework (Ma *et al.*, 2006) shows that populations of neurons can implicitly represent probability distributions over the stimuli. To see how, consider a stimulus s evoking activity in a population of N neurons, denoted as

$$\mathbf{r} = (r_1, \ldots, r_N).$$

In the standard view of neural population decoding, the identical stimulus s would be presented over and over again, generating an estimate of the distribution $p(\mathbf{r}|s)$. In contrast, by a Bayesian argument, the PPC framework yields an estimate of the entire probability distribution (known as posterior distribution) over the stimulus range:

$$p(s|\mathbf{r}) \propto p(\mathbf{r}|s)p(s), \tag{28}$$

which is valid because $p(\mathbf{r})$ is merely a normalization factor that ensures that $p(s|\mathbf{r})$ is a proper probability distribution over s. Note that (28) holds even if the prior $p(s)$ is flat (uniform). As formulated in Ma *et al.* (2008), it is necessary to consider a population rather than a single neuron because the set of distributions that can be encoded by a single neuron is very limited. For example, a normal distribution already has two independent parameters, mean and variance, and thus requires at least two neurons to represent it.

Obviously, for the Bayesian PPC approach to work, the distribution of $p(\mathbf{r}|s)$ must be known. The simplest case, which we follow below, is to assume that it can be described as a set of independent Poisson processes. Given that neurons do not exactly follow Poisson and are correlated, more general approaches have been proposed, in particular so-called Poisson-like variability, that is, the exponential family with linear sufficient statistics (Ma *et al.*, 2006).

Following the independent Poisson case, the probability for spike count r_i of neuron i in response to a stimulus s is

$$p(r|s) = \frac{\mathrm{e}^{-\lambda_i}\lambda_i^r}{r!},$$

with $\lambda_i > 0$. Mean spike count is assumed to be $\lambda_i = gf_i(s)$ where g is an overall scaling factor (the gain), $f_i(s)$ is the tuning curve for neuron i with Gaussian shape. Thus, the likelihood function $p(\mathbf{r}|s)$ (as function of s) equals

$$\begin{aligned} p(\mathbf{r}|s) = p(r_1, r_2, \ldots, r_N|s) &= p(r_1|s) \times p(r_2|s) \ldots \times p(r_N|s) \\ &= \prod_{i=1}^{N} \frac{\mathrm{e}^{-gf_i(s)}(gf_i(s))^{r_i}}{r_i!}. \end{aligned}$$

From (28), with a uniform prior, the posterior distribution obtains,

$$p(s|\mathbf{r}) \propto \exp \sum_{i=1}^{N} (-gf_i(s) + r_i \log f_i(s)).$$

This is not a probability distribution in the frequentist sense as we are considering only a single trial. It provides not only the most likely value of the stimulus, but also its uncertainty via the width of the distribution. A higher gain implies narrower posteriors and less uncertainties.

Interestingly, the PPC framework allows for a Bayes optimal combination of crossmodal cues by a simple linear summation of neural population activity. We assume that each cue (visual, auditory) is represented by N neurons, with independent activity patterns $\mathbf{r}_A$ and $\mathbf{r}_V$ and identical tuning curves $f_i(s)$, differing only in their gains, so that the mean activities are $g_A f_i(s)$ and $g_V f_i(s)$. By conditional independence,

$$p(s|\mathbf{r}_A, \mathbf{r}_V) \propto p(\mathbf{r}_A, \mathbf{r}_V|s) = p(\mathbf{r}_A|s)p(\mathbf{r}_V|s)$$
$$p(s|\mathbf{r}_A, \mathbf{r}_V) \propto \exp \sum_{i=1}^{N} [-(g_A + g_V)f_i(s) + (r_{Ai} + r_{Vi}) \log f_i(s)].$$

We want a single multisensory population: What operation has to be performed on $\mathbf{r}_A$ and $\mathbf{r}_V$ such that the resulting multisensory distribution encodes the optimal posterior distribution, that is, not losing any information about s? We construct a new population pattern of activity, $\mathbf{r}_{AV}$, by summing the activities of corresponding pairs of neurons in the visual and auditory populations:

$$\mathbf{r}_{AV} = \mathbf{r}_A + \mathbf{r}_V. \tag{29}$$

The output pattern $\mathbf{r}_{AV}$ will still obey independent Poisson variability across many trials, as the sum of two Poisson processes is again Poisson. The mean activity of the ith neuron in the output population in response to s is $(g_A + g_V)f_i(s)$. Therefore, it encodes a posterior distribution that is given by

$$p(s|\mathbf{r}_{AV}) \propto \exp \sum_{i=1}^{N} [-(g_A + g_V)f_i(s) + (r_{Ai} + r_{Vi}) \log f_i(s)].$$

This is identical to the above one: Adding independent Poisson population patterns of activity implements a multiplication of the probability distributions over the stimulus that are encoded in those patterns.

As mentioned above, when neural variability is in the Poisson-like family with correlation, optimal cue integration—a multiplicative operation at the level of the posteriors—is still realized through a simple linear combination of population responses irrespective of the shapes of the tuning curves or the covariance matrices (see Ma *et al.*, 2006). The PPC framework was used in analyzing recordings from area MSTd in monkeys in a visual–vestibular heading task (Fetsch *et al.*, 2011). Although cue reliability varied randomly from trial to trial, the monkeys appropriately placed greater weight on the more reliable cue, and population decoding of neural responses in the MSTd closely predicted behavioral cue weighting, including modest deviations from optimality.

Note the obvious analogy between the additivity of the activity patterns in Eqn (29) and the minimum-variance linear combination rule of the standard Gaussian cue combination model of Eqn (6). Given that the variances are automatically taken into account appropriately through the interplay of neural variability (Poisson-like) and network operations (linear combination) in a single trial, this has been interpreted as guaranteeing that knowledge of the variances (reliabilities) required for optimal behavior is actually represented in the brain (Ma & Pouget, 2008). Nevertheless, Fetsch *et al.* (2013) argue convincingly that three different uses of the term "weights" need to be distinguished: (i) Synaptic input weights from primary sensory neurons onto a multisensory neuron reflecting the number and/or efficacy of synaptic connections associated with each modality are different from what (ii) is measured in most single-unit studies of multisensory integration and, finally (iii), the relationship between those neural weights and the perceptual weights measured in psychophysical tasks is mostly unknown yet. In this context, Ma & Rahmati (2013) presented a PPC implementation of the Bayesian causal inference model but concluded that the resulting architecture is unrealistic as they needed a large number of specific combinations of the constituent elements to realize the optimal decision rule.

Divisive normalization

Divisive normalization is often seen as a near-universal feature of neural computation across many brain areas involving vision, audition, olfaction, spatial attention, and higher cognitive processes (for a recent review, see, Carandini & Heeger, 2012). Normalization computes a ratio between the response of an individual neuron and the summed activity of a pool of neurons; this equation specifies how the normalized response R of a neuron depends on its inputs (which are not normalized):

$$R = \frac{E^n}{\alpha^n + \left(\frac{1}{N}\right) \sum_{j=1}^{N} E_j^n}. \tag{30}$$

The denominator is a constant α plus the normalization factor, which is the sum of a large number of inputs, the normalization pool. The constants α and n are parameters that are typically fit to empirical measurements. α prevents division by zero and determines how responses saturate with increasing input, and n is an exponent that amplifies the individual inputs.

Ohshiro *et al.* (2011) developed a specific divisive normalization model for multisensory integration. The unisensory inputs to each multisensory neuron increase monotonically, but sublinearly, with stimulus intensity, modeling response saturation in the sensory inputs (e.g., by synaptic depression). Each multisensory neuron performs a weighted linear sum of its unisensory inputs with weights d_1 and d_2, named modality dominance weights,

$$E = d_1 I_1(x_0, y_0) + d_2 I_2(x_0, y_0). \tag{31}$$

Here, $I_1(x_0, y_0)$ and, $I_2(x_0, y_0)$ represent the two unisensory inputs, indexed by spatial location of the receptive fields. The modality dominance weights are fixed for each multisensory neuron, but different neurons have different combinations of d_1 and d_2 to simulate various degrees of dominance of one sensory modality. The final response of a multisensory neuron is then determined by Eqn (30). Ohshiro and colleagues could then show that this normalization model can account for the classic empirical integration rules observed in dSC: inverse effectiveness, spatial principle, and also multisensory suppression in unisensory neurons.

Models for multisensory responses in dSC

A number of computational models have been developed that specifically target the empirical findings on dSC multisensory neurons (Alvarado *et al.*, 2008; Rowland & Stein, 2014). They go beyond models like the one by Ohshiro *et al.* (2011) (i) in representing the temporal profile of multisensory enhancement or suppression (Rowland *et al.*, 2007), (ii) in explicitly taking into account the finding that multisensory integration in the SC is mediated by descending inputs from association cortex (Jiang *et al.*, 2001), and (iii) in addressing questions about the emergence and maturation of multisensory integration in dependence of the environment (Stein *et al.*, 2014). In a recent computational model, the continuous-time multisensory model (CTMM) by Miller *et al.* (2017), the authors show that real-time integration and delayed, calibrating inhibition are sufficient to predict the individual moment-by-moment multisensory response given only knowledge of the associated unisensory responses. Some aspects of CTMM are reminiscent of the MCD correlation detector by Parise & Ernst (2016), and it might be of interest to investigate this further.

Another approach to modeling multisensory integration in the SC are the neural network models by Cuppini, Ursino, Magosso, and

colleagues. Typical assumptions of these models are that neurons of the two modality have different receptive fields (the visual ones more accurate in the spatial domain, the auditory ones more accurate in the temporal domain) and that the crossmodal synapses mainly link neurons with proximal spatial preference. With these models, they are able to simulate some typical audiovisual illusions (such as the ventriloquism effect) including aftereffects. The basic hypothetical assumption of the model developed in Cuppini *et al.* (2010) is that, in the cat, the operation of the SC is under normal circumstances almost entirely controlled by the sensory inputs derived from cortical areas AES and rLS. The neural net could account for multisensory enhancement and suppression, inverse effectiveness, the effects of selective cortical deactivation, NMDA blockade, and the differing responses and underlying computations that characterize responses to pairs of spatially disparate crossmodal and within-modal stimuli (for details, see Cuppini *et al.*, 2010). In an effort to construct the neural underpinnings of Bayesian inference, a recent model by the same authors (Ursino *et al.*, 2017) shows how the PPC approach, together with Hebbian reinforcement and a decay term during training, can shrink the receptive fields and can encode the unisensory likelihood functions.

Conclusion

We aimed at presenting a snapshot of (mostly) recent modeling efforts to understand multisensory integration. Although quite a number of models connected with experimental paradigms requiring the processing of crossmodal information have been selected, it is obvious that these paradigms represent only a fraction of the multisensory contexts that humans and animals experience in a normal environment, that is, outside the laboratory. Moreover, completeness is not claimed here, neither with respect to experimental paradigms nor formal modeling approaches. One area which would perhaps deserve a separate review is audiovisual speech integration treating; for example, models developed for the McGurk–McDonald effect (e.g., Magnotti & Beauchamp, 2017). Another area comprises multisensory aspects of development and of specific clinical populations (e.g., autism spectrum) that could give an opportunity to probe how the different models would be able to deal with the findings. There was also no space to deal with many models at sufficient detail and, in particular, to discuss their performance vis-à-vis empirical data. Multisensory modeling is a very dynamic area of research, so it may be too early for sweeping conclusions. Nevertheless, let us conclude with some observations, listed without a particular order of importance.

First, quantitative measures of the amount of multisensory integration, like CRE in Eqn (1) and its variants, are still being used widely, providing a means to compare levels of integration across different experimental paradigms and variables (spike numbers, reaction times, detection probability). It remains to be seen whether the modification of CRE based on probability summation (proposed in Colonius & Diederich, 2017) will find broad acceptance. Second, the "simple integration rules" of the effects of space, time, and energy ("inverse effectiveness") hold up in many experiments, but their role in modeling diminishes the more complex the tasks become, like in some of the judgment paradigms. Third, the progress in developing empirically testable, quantitative models over the last 15 to 20 years is remarkable; unfortunately, model testing is often limited to comparing some variants of one and the same conceptual approach. Progress toward a unifying theory would greatly benefit from devising critical experiments that allow to distinguish between broader classes of models, although this may be quite challenging and not always feasible. Fourth, most multisensory integration models are based on very general mechanisms, such as probability summation, Bayesian cue combination, or neural networks, which have been in use in many other areas of computational neuroscience and behavioral modeling as well. As such, this may not be surprising; still, one may hypothesize that the development of a unifying theoretical approach to multisensory integration, narrowing the gap between neural and behavioral approaches, may benefit from identifying concepts or mechanisms that uniquely characterize the integration of information from different sensory systems.

Conflict of interest

Authors declare no conflict of interest.

Acknowledgements

Constructive and very informative comments by two reviewers are gratefully acknowledged. Thanks also to Maike Tahden and Anja Gieseler for checking the manuscript. Authors take full responsibility for any remaining faults and expressions of opinions.

Data accessibility

Not applicable.

Author Contributions

Both authors contributed equally to the writing of the review paper.

Funding information

Supported by DFG (German Science Foundation SFB/TRR-31 (Project B4, HC), DFG Cluster of Excellence EXC 1077/1 Hearing4all (HC) and DFG Grant DI 506/12-1 (AD).

References

Alais, D. & Burr, D. (2004) The ventriloquist effect results from near optimal crossmodal integration. *Curr. Biol.*, **14**, 257–262.
Alvarado, J., Rowland, B., Stanford, T. & Stein, B. (2008) A neural network model of multisensory integration also accounts for unisensory integration in superior colliculus. *Brain Res.*, **1242**, 13–23.
Beauchamp, M.S. (2005) Statistical criteria in fMRI studies of multisensory integration. *Neuroinformatics*, **3**, 93–113.
Bhattacharya, R.N. & Waymire, E. (1990). *Stochastic Processes with Applications*. Wiley, New York.
Borst, A. & Euler, T. (2011) Seeing things in motion: models, circuits, and mechanisms. *Neuron*, **71**, 974–994.
Bremner, A., Lewkowicz, D. & Spence, C. (Eds) (2012). *Multisensory Development*. Oxford University Press, Oxford, UK.
Bresciani, J.-P., Dammeier, F. & Ernst, M. (2006) Vision and touch are automatically integrated for the perception of sequences of events. *J. Vision*, **6**, 554–564.
Burr, D., Silva, O., Cichini, G., Banks, M.S. & Morrone, M. (2009) Temporal mechanisms of multimodal binding. *Proc. R. Soc. B*, **276**, 1761–1769.
Calvert, G., Spence, C. & Stein, B. (Eds) (2004). *Handbook of Multisensory Processes: Modeling the Time Course of Multisensory Perception in Manual and Saccadic Reponses*. MIT Press, Cambridge, MA.
Carandini, M. & Heeger, D. (2012) Normalization as a canonical neural computation. *Nat. Rev. Neurosci.*, **13**, 51–62.
Chen, L. & Vroomen, J. (2013) Intersensory binding across space and time: a tutorial review. *Atten. Percept. Psycho.*, **75**, 790–811.
Colonius, H. (1990) Possibly dependent probability summation of reaction time. *J. Math. Psychol.*, **34**, 253–275.
Colonius, H. (2015) Behavioral measures of multisensory integration: bounds on bimodal detection probability. *Brain Topogr.*, **28**, 1–4.

Colonius, H. (2016) An invitation to coupling and copulas, with applications to multisensory modeling. *J. Math. Psychol.*, **74**, 2–10.
Colonius, H. & Diederich, A. (2004) Multisensory interaction in saccadic reaction time: a time-window-of-integration model. *J. Cognitive Neurosci.*, **16**, 1000–1009.
Colonius, H. & Diederich, A. (2006) The race model inequality: interpreting a geometric measure of the amount of violation. *Psychol. Rev.*, **113**, 148–154.
Colonius, H. & Diederich, A. (2017) Measuring multisensory integration: from reaction times to spike counts. *Sci. Rep.*, **7**, 3023.
Colonius, H., Wolff, F. & Diederich, A. (2017) Trimodal race model inequalities in multisensory integration: I. Basics. *Front. Psychol.*, **8**, 1141.
Cuppini, C., Ursino, M., Magosso, E., Rowland, B. & Stein, B. (2010) An emergent model of multisensory integration in superior colliculus neurons. *Front. Integr. Neurosci.*, https://doi.org/10.3389/fnint.2010.00006. [Epub ahead of print.]
Debats, N., Ernst, M. & Heuer, H. (2017) Perceptual attraction in tool use: evidence for a reliability-based weighting mechanism. *J. Neurophysiol.*, **117**, 1569–1580.
Diederich, A. (1992a). *Intersensory Facilitation: Race, Superposition, and Diffusion Models for Reaction Time to Multiple Stimuli*. Peter Lang, Frankfurt.
Diederich, A. (1992b) Probability inequalities for testing separate activation models of divided attention. *Perception & Psychophysics*, **52**, 714–716.
Diederich, A. (1995) Intersensory facilitation of reaction time: evaluation of counter and diffusion coactivation models. *J. Math. Psychol.*, **39**, 197–215.
Diederich, A. (2003) MDFT account of decision making under time pressure. *Psychon. B. Rev.*, **10**, 157–166.
Diederich, A. & Colonius, H. (1991) A further test of the superposition model for the redundant signals effect in bimodal detection. *Perception & Psychophysics*, **50**, 83.
Diederich, A. & Colonius, H. (2004a) Bimodal and trimodal multisensory enhancement: effects of stimulus onset and intensity on reaction time. *Percept. Psychophys.*, **66**, 1388–1404.
Diederich, A. & Colonius, H. (2004b). Modeling the time course of multisensory interaction in manual and saccadic responses. In Calvert, G., Spence, C. & Stein, B. (Eds), *Handbook of Multisensory Processes*. MIT Press, Cambridge, MA, pp. 395–408.
Diederich, A. & Colonius, H. (2007a) Modeling spatial effects in visualtactile reaction time. *Percept. Psychophys.*, **69**, 56–67.
Diederich, A. & Colonius, H. (2007b) Why two "distractors" are better then one: modeling the effect on non-target auditory and tactile stimuli on visual saccadic reaction time. *Exp. Brain Res.*, **179**, 43–54.
Diederich, A. & Colonius, H. (2008a) Crossmodal interaction in saccadic reaction time: separating multisensory from warning effects in the time window of integration model. *Exp. Brain Res.*, **186**, 1–22.
Diederich, A. & Colonius, H. (2008b) When a high-intensity "distractor" is better then a low-intensity one: modeling the effect of an auditory or tactile nontarget stimulus on visual saccadic reaction time. *Brain Res.*, **1242**, 219–230.
Diederich, A. & Colonius, H. (2015) The time window of multisensory integration: relating reaction times and judgments of temporal order. *Psychol. Rev.*, **122**, 232–241.
Drugowitsch, J., DeAngelis, G., Klier, E.M., Angelaki, D. & Pouget, A. (2014) Optimal multisensory decision-making in a reaction-time task. *elife*, **3**, e03005.
Ernst, M. (2012). Optimal multisensory integration: assumptions and limits. In Stein, B. (Ed), *The New Handbook of Multisensory Processes*. MIT Press, Cambridge, MA, pp. 527–544.
Ernst, M. & Banks, M.S. (2002) Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, **415**, 429–433.
Ernst, M. & Di Luca, M. (2011). Multisensory perception: from integration to remapping. In Trommershäuser, J., Körding, K. & Landy, M. (Eds), *Sensory Cue Integration, chap 12*. Oxford University Press, Oxford, UK, pp. 224–250.
Fetsch, C., Pouget, A., DeAngelis, G. & Angelaki, D. (2011) Neural correlates of reliability-based cue weighting during multisensory integration. *Nat. Neurosci.*, **15**, 146–154.
Fetsch, C., DeAngelis, G. & Angelaki, D. (2013) Bridging the gap between theories of sensory cue integration and the physiology of multisensory neurons. *Nat. Rev. Neurosci.*, **14**, 429–442.
Fisher, Y., Sillies, M. & Clandinin, T. (2015) Orientation selectivity sharpens motion detection in Drosophila. *Neuron*, **88**, 390–402.
Fujisaki, W. & Nishida, S. (2007) Feature-based processing of audio-visual synchrony perception revealed by random pulse trains. *Vision. Res.*, **47**, 1075–1093.
Fujisaki, W., Shimojo, S., Kashino, M. & Nishida, S. (2004) Recalibration of audiovisual simultaneity. *Nat. Neurosci.*, **7**, 773–778.
García-Pérez, M. & Alcalá-Quintana, R. (2012) On the discrepant results in synchrony judgment and temporal-order judgment tasks: a quantitative model. *Psychon. B. Rev.*, **19**, 820–846.
Gondan, M. & Minakata, K. (2016) A tutorial on testing the race model inequality. *Atten. Percept. Psycho.*, **78**, 723–735.
Gondan, M., Lange, K., Rösler, F. & Röder, B. (2004) The redundant target effect is affected by modality switch costs. *Psychon. B. Rev.*, **11**, 307–313.
Green, D. & Swets, J. (1966). *Signal Detection Theory and Psychophysics*. Wiley, New York.
Harrar, V., Harris, L. & Spence, C. (2016) Multisensory integration is independent of perceived simultaneity. *Exp. Brain Res.*, **235**, 763–775.
Hassenstein, V. & Reichardt, W. (1956) System theoretical analysis of time, sequence and sign analysis of the motion perception of the snout-beetle chlorophanus. *Z. Naturforsch.*, **11**, 513–524 (in German).
Jeffress, L. (1948) A place theory of sound localization. *J. Comp. Physiol. Psych.*, **41**, 35–39.
Jiang, W., Wallace, M., Jiang, H., Vaughan, J. & Stein, B. (2001) Two cortical areas mediate multisensory integration in superior colliculus neurons. *J. Neurophysiol.*, **85**, 506–522.
Juan, C., Cappe, C., Alric, B., Roby, B., Gilardeau, S., Barone, P. & Girard, P. (2017) The variability of multisensory processes of natural stimuli in human and non-human primates in a detection task. *PLoS One*, **12**, E0172480.
Kayser, C. & Shams, L. (2015) Multisensory causal inference in the brain. *PLoS Biol.*, **13**, e1002075.
Knill, D. & Richards, W. (Eds) (1996). *Perception as Bayesian Inference*. Cambridge University Press, Cambridge, UK.
Körding, K., Beierholm, U., Ma, W., Quartz, S., Tenenbaum, J. & Shams, L. (2007) Causal inference in multisensory perception. *PLoS One*, **2**, e943.
Luce, R. (1986). *Response Times: Their Role in Inferring Elementary Mental Organization*. Oxford University Press, New York, NY.
Ma, W. & Pouget, A. (2008) Linking neurons to behavior in multisensory perception: a computational review. *Brain Res.*, **1242**, 4–12.
Ma, W. & Rahmati, M. (2013) Towards a neural implementation of causal inference in cue combination. *Multisens. Res.*, **26**, 159–176.
Ma, W., Beck, J., Latham, P. & Pouget, A. (2006) Bayesian inference with probabilistic population codes. *Nat. Neurosci.*, **9**, 1432–1438.
Ma, W., Beck, J. & Pouget, A. (2008) Spiking networks for Bayesian inference and choice. *Curr. Opin. Neurobiol.*, **18**, 217–222.
Magnotti, J. & Beauchamp, M. (2017) A causal inference model explains perception of the mcgurk effect and other incongruent audiovisual speech. *PLoS Comput. Biol.*, **13**, e1005229.
Mari, D. & Kotz, S. (2001). *Correlation and Dependence*. Imperial College Press, London.
McGurk, H. & MacDonald, J. (1976) Hearing lips and seeing voices. *Nature*, **264**, 746–748.
Mégevand, P., Molholm, S., Nayak, A. & Foxe, J.J. (2013) Recalibration of the multisensory temporal window of integration results from changing task demands. *PLoS One*, **8**, e71608.
Mendonca, C., Escher, A., van de Par, S. & Colonius, H. (2015) Predicting auditory space calibration from recent multisensory experience. *Exp. Brain Res.*, **233**, 1983–1991.
Meredith, M. (2002) On the neuonal basis for mulitsensory convergence: a brief overview. *Cognitive Brain Res.*, **14**, 31–40.
Meredith, M. & Stein, B. (1983) Interactions among converging sensory inputs in the superior colliculus. *Science*, **221**, 389–391.
Miller, J. (1982) Divided attention: evidence for coactivation with redundant signals. *Cognitive Psychol.*, **14**, 247–279.
Miller, J. (1986) Timeocurs of coactivation in bimodal divided atttention. *Perception & Psychophysics*, **40**, 331–343.
Miller, J. (2016) Statistical facilitation and the redundant signals effect: what are race and coactivation models? *Atten. Percept. Psycho.*, **78**, 516–519.
Miller, R., Pluta, S., Stein, B. & Rowland, B. (2015) Relative unisensory strength and timing predict their multisensory product. *J. Neurosci.*, **35**, 5213–5220.
Miller, R., Stein, B. & Rowland, B. (2017) Multisensory integration uses a real time unisensory-multisensory transform. *J. Neurosci.*, **37**, 5183–5193.
Mitrinović, D. (1970). *Analytic Inequalities, vol. 165 of Die Grundlehren der mathematischen Wissenschaften*. Springer-Verlag, Berlin.
Miyazaki, M., Yamamoto, S., Uchida, S. & Kitazawa, S. (2006) Bayesian calibration of simultaneity in tactile temporal order judgment. *Nat. Neurosci.*, **9**, 875–877.

Murray, M. & Wallace, M. (Eds) (2012). *The Neural Bases of Multisensory Processes. Frontiers in Neuroscience*. CRC Press, Boca Rato, FL.
Naumer, M. & Kayser, J. (Eds) (2010). *Multisensory Object Perception in the Primate Brain*. Springer- Verlag, New York, NY.
Odegaard, B., Wozny, D. & Shams, L. (2015) Biases in visual, auditory, and audiovisual perception of space. *PLoS Comput. Biol.*, **11**, e1004649. https://doi.org/10.1371/journal.pcbi.1004649. [Epub ahead of print].
Ohshiro, T., Angelaki, D. & DeAngelis, G. (2011) A normalization model of multisensory integration. *Nat. Neurosci.*, **14**, 775–782.
Ohzawa, I. (1998) Mechanisms of stereoscopic vision: the disparity energy model. *Curr. Opin. Neurobiol.*, **8**, 509–515.
Oruç, I., Maloney, L. & Landy, M.S. (2003) Weighted linear cue combination with possibly correlated error. *Vision*, **43**, 2451–2468.
Parise, C. & Ernst, M. (2016) Correlation detection as a general mechanism for multisensory integration. *Nat. Commun.*, **7**, https://doi.org/10.1038/ncomms11543. [Epub ahead of print].
Parise, C. & Spence, C. (2009) 'When birds of a feather ock together': synesthetic correspondences modulate audiovisual integration in non-synesthetes. *PLoS One*, **4**(5), e5664.
Parise, C., Spence, C. & Ernst, M. (2012) When correlation implies causation in multisensory integration. *Curr. Biol.*, **22**, 1–4.
Parise, C., Harrar, V., Ernst, M. & Spence, C. (2013) Cross-correlation between auditory and visual signals promotes multisensory integration. *Multisens. Res.*, **26**, 307–316.
Perrault Jr, T., Vaughan, J., Stein, B. & Wallace, M. (2005) Superior colliculus neurons use distinct operational modes in the integration of multisensory stimuli. *J. Neurophysiol.*, **93**, 2575–2586.
Populin, L. & Yin, T. (2002) Bimodal interactions in the superior colliculus of the behaving cat. *J. Neurosci.*, **22**, 2826–2834.
Pouget, A., Deneve, S. & Duhamel, J. (2002) A computational perspective on the neural basis of multisensory spatial representations. *Nat. Rev. Neurosci.*, **3**, 741–747.
Raab, D. (1962) Statistical facilitation of simple reaction time. *T. New York Acad. Sci.*, **24**, 574–590.
Ricciardi, L. (1977). *Diffusion Processes and Related Topics in Biology*. Springer Verlag, Berlin.
Rohe, T. & Noppeney, U. (2015a) Cortical hierarchies perform bayesian causal inference in multisensory perception. *PLoS Biol.*, **13**, e1002073.
Rohe, T. & Noppeney, U. (2015b) Sensory reliability shapes perceptual inference via two mechanisms. *J. Vision*, **15**, 1–16.
Rosas, P. & Wichmann, F. (2011). Cue combination: beyond optimality. In Trommershäuser, J., Körding, K. & Landy, M. (Eds), *Sensory Cue Integration, chap 8*. Oxford University Press, Oxford, pp. 144–152.
Rowland, B. & Stein, B. (2014) A model of the temporal dynamics of multisensory enhancement. *Neurosci. Biobehav. R.*, **41**, 78–84.
Rowland, B., Quessy, S., Stanford, T. & Stein, B. (2007) Multisensory integration shortens physiological response latencies. *J. Neurosci.*, **27**, 5879–5884.
van Santen, J. & Sperling, G. (1985) Elaborated Reichardt detectors. *J. Opt. Soc. Am. A*, **2**, 300–321.
Sato, Y. & Aihara, K. (2011) A Bayesian model of sensory adaptation. *PLoS One*, **6**, e19377.
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.
Schwarz, W. (1989) A new model to explain the redundantsignals effect. *Perception & Psychophysics*, **46**, 490–500.
Schwarz, W. (1994) Diffusion, superposition, and the redundant-targets effect. *J. Math. Psychol.*, **38**, 504–520.
Shams, L. & Beierholm, U. (2010) Causal inference in perception. *Trends Cogn. Sci.*, **14**, 425–432.
Shaw, M. (1982) Attending to multiple sources of information: I: The integration of information in decision making. *Cognitive Psychol.*, **14**, 353–409.
Smith, P. & van Zandt, T. (2000) Timedependent poisson counter models of response latency in simple judgment. *Brit. J. Math. Stat. Psy.*, **53**, 293–315.
Stein, B. (Ed) (2012). *The New Handbook of Multisensory Processes*. MIT Press, Cambridge, MA.
Stein, B. & Meredith, M. (1993). *The Merging of the Senses*. MIT Press, Cambridge, MA.
Stein, B., Stanford, T., Ramachandran, R., Perrault Jr, T. & Rowland, B. (2009) Challenges in quantifying multisensory integration: alternative criteria, models, and inverse effectiveness. *Exp. Brain Res.*, **198**, 113–126.
Stein, B., Burr, D., Constantinidis, C., Laurienti, P., Meredith, M., Perrault Jr, T. , Ramachandran, R., Roeder, B. *et al.* (2010) Semantic confusion regarding the development of multisensory integration: a practical solution. *Eur. J. Neurosci.*, **31**, 1713–1720.
Stein, B., Stanford, T. & Rowland, B. (2014) Development of multisensory integration from the perspective of the individual neuron. *Nat. Rev. Neurosci.*, **15**, 520–535.
Todd, J. (1912). *Reaction to multiple stimuli. Archives of Psychology No. 25. Columbia Contributions to Philosophy and Psychology, XXI*. The Science Press, New York.
Townsend, J. & Nozawa, G. (1995) Spatiotemporal properties of elementary perception: an investigation of parallel, serial, and coactive theories. *J. Math. Psychol.*, **39**, 321–359.
Tuckwell, H. (1995). *Elementary Applications of Probability Theory - With an Introduction to Stochastic Differential Equations*, 2nd edn. Chapman and Hall, London.
Ursino, M., Cuppini, C. & Magosso, E. (2017) Multisensory bayesian inference depends on synapse maturation during training: theoretical analysis and neural modeling implementation. *Neural Comput.*, **29**, 735–782.
Vroomen, J. & Keetels, M. (2010) Perception of intersensory synchrony: a tutorial review. *Atten. Percept. Psycho.*, **72**, 871–884.
Vulkan, N. (2000) An economist's perspective on probability matching. *J. Econ. Surv.*, **14**, 101–118.
de Winkel, K., Katliar, M. & Bülthoff, H. (2015) Forced fusion in multisensory heading estimation. *PLoS One*, **10**, e0127104.
de Winkel, K., Katliar, M. & Bülthoff, H. (2017) Causal inference in multisensory heading estimation. *PLoS One*, **12**, e016967.
Wozny, D. & Shams, L. (2011) Recalibration of auditory space following milliseconds of cross-modal discrepancy. *J. Neurosci.*, **31**, 4607–4612.
Wozny, D., Beierholm, U. & Shams, L. (2010) Probability matching as a computational strategy used in perception. *PLoS Comput. Biol.*, **6**, e1000871.
Yamamoto, S., Miyazaki, M., Iwano, T. & Kitazawa, S. (2012) Bayesian calibration of simultaneity in audiovisual temporal order judgments. *PLoS One*, **7**, e40379.
Zaidel, A., Ma, W. & Angelaki, D. (2013) Supervised calibration relies on the multisensory percept. *Neuron*, **80**, 1544–1557.
van Zandt, T., Colonius, H. & Proctor, R. (2000) A comparison of two response time models applied to perceptual matching. *Psychon. B. Rev.*, **7**, 208–256.