

FORMAL MODELS AND QUANTITATIVE MEASURES OF
MULTISENSORY INTEGRATION:
A SELECTIVE OVERVIEW

HANS COLONIUS

Department of Psychology
Carl von Ossietzky Universität Oldenburg
and
Department of Psychological Sciences
Purdue University

ADELE DIEDERICH

Life Sciences and Chemistry
Jacobs University Bremen
and
Department of Psychological Sciences
Purdue University

ABSTRACT. This review presents a selective overview of formal modeling approaches to multisensory integration (MI). First, measures of MI in different experimental paradigms are discussed. The focus is on different approaches to modeling (crossmodal) cue combination. Models for psychophysical estimation and judgment paradigms including the Bayesian Causal Inference model are presented. Reaction time modeling approaches like race model, coactivation models, and the time window of integration model are discussed. The final section considers computational neuronal models including probabilistic population code models, divisive normalization, and neural network models for multisensory responses in superior colliculus.

Key words: Bayesian causal inference, race model, time window of integration, probabilistic population codes, diffusion model

E-mail addresses: `hans.colonius@uol.de`, `a.diederich@jacobs-university.de`.

Date: October 17, 2017.

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).

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 ([9, 90, 5, 66, 67]). Early studies by Todd in 1912 [95], measuring reaction time to 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 [55] in speech perception or, at the neurophysiological level, the identification of “multi-sensory” neurons in several brain areas [59] 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 data base 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 is 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 (i) serves to clarify which phenomena are to be explained; (ii) it renders more precise the 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; and (iii) it

enables 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, subsequently, 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 [74].

Finally, it seems difficult to identify modeling concepts that are specific to multisensory processing as such. Most current multisensory integration models are based on very general mechanisms, like probability summation, Bayesian cue combination or neural networks that have been in use in many other areas of computational neuroscience and behavior as well.

MEASURING MULTISENSORY INTEGRATION

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 [91], the sixteen authors, working in areas from basic neuroscience to psychology, identified a wide-spread 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

Definition 1 (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 “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 and the modality-specific responses, since it directly leads to the original quantitative index of multisensory integration for the number of spikes emitted in a fixed time interval after stimulation [59, 92].

Definition 2.

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

where CM is the mean (absolute) number of spikes in response to the cross-modal stimulus and 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 [76]. Prominently, in the “additive model”, term SM_{\max} in Equation (1) is replaced by the sum of the unisensory responses [77]. The additive version has raised some controversy because, under some modeling assumptions, an additive combination of cross-modal inputs yields a prediction of optimal multisensory integration [78]. 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 [93]. At this point it is important to point out that our discussion here is necessarily incomplete since the important case of *multisensory inhibition* is omitted due to limits of space.

While CRE and its variations have obvious value as descriptive measurement tools, we see two important shortcomings. First, the only information they are based on are the mean responses although other parameters, like variance and temporal alignment of stimuli, may also carry important information. For example, in a recent study [63], Rowland 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),

$$(2) \quad \text{UI} = \frac{V - A}{V + A},$$

plays a decisive role in predicting the strength of multisensory integration. Specifically, their findings suggest the “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). Second, CRE falls short of providing any insight into the underlying mechanism of integration.

We recently suggested an alternative measure :

$$(3) \quad \text{CRE}^- = \frac{\text{CM} - \text{SM}^-}{\text{SM}^-} \times 100.$$

Here, SM^- replaces SM ; it refers to the maximal mean achievable by combining the unisensory data under a probability summation rule with negative dependence. CRE^- is sensitive to the entire distribution of spikes and also more conservative in never predicting more multisensory integration than CRE . In a sample of spike data, we recently showed that this new measure will reduce the proportion of neurons so far labeled as “multisensory” [16].

Modifying Definition 1 to become applicable to measures other than spike numbers is straightforward. For reaction times, multisensory integration typically manifest itself by speeded responses to crossmodal stimuli [95, 19]. Then CRE (Equation 1) becomes [24],

$$(4) \quad \text{CRE}_{RT} = \frac{\min\{\mathbf{ERT}_V, \mathbf{ERT}_{A+\tau}\} - \mathbf{ERT}_{VA}}{\min\{\mathbf{ERT}_V, \mathbf{ERT}_{A+\tau}\}} \times 100.$$

where \mathbf{ERT}_{VA} is the mean RT to the crossmodal stimulus (taken as combination of visual and acoustic stimuli here) and $\min\{\mathbf{ERT}_V, \mathbf{ERT}_{A+\tau}\}$ is the faster of the unisensory mean RTs to the visual stimulus and the acoustic stimulus presented at stimulus onset asynchrony (SOA) τ ($-\infty < \tau < +\infty$). Thus, CRE_{RT} expresses multisensory enhancement as a proportional reduction of the faster unisensory response by the crossmodal response. For example, $\text{CRE}_{RT} = 10$ means that response 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

$$(5) \quad \text{CRE}_{DP} = \frac{p_{VA} - \max\{p_V, p_A\}}{\max\{p_V, p_A\}},$$

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

All these versions of CRE for different paradigms are subject to the same two shortcomings mentioned above. Alternative versions, based on probability summation analogous to measure CRE^- in Equation (3), can be developed for them (for an example, see [12]). In summary, the issue of exactly how to measure the amount of multisensory integration has been under debate for a long time and is likely to continue [93, 76, 2].

MODELING MULTISENSORY INTEGRATION: WHAT NEEDS TO BE MODELED?

A defining feature of multisensory integration is the distinction between a unisensory 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 spatio-temporal 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, e.g. \mathcal{V} and \mathcal{A} , are related to responses to signals from the crossmodal context, e.g. \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 sub-populations, 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 effect where the crossmodal 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 “spatio-temporal rule” holds that stimuli presented in spatio-temporal proximity have a higher probability of being integrated (the “spatio-temporal window” of integration, see [58]). Second, the “inverse-effectiveness rule” predicts that the amount of integration increases when the intensity level of the unisensory 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 “spatio-temporal 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 integration 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, e.g., 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 like 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 [40, 88]. In a similar vein, introducing Bayesian concepts into

perceptual modeling has a long tradition, see [47] 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, e.g. the size of an object to be estimated by an observer via two sensory modalities, e.g. 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, [33] 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:

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

The following optimal solution is adapted from [71] :

Proposition 1. The variance of $w_V S_V + w_H S_H$ is minimized when the weights are

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

yielding

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

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

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

Remarkably, the proof (see appendix for 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 appreciated when the linear ‘‘Gaussian cue combination’’ model is probed on a set of data. On the other hand, when nonlinear cue combinations are not excluded a-priori, the Gaussian assumption is sufficient to prove the minimum variance unbiased property of the weighted linear combination. Since in this case the solution is also the maximum likelihood estimate, the model is often called *MLE model*.

The case of *correlated* cue distributions and negative weights is discussed in [71] 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 and colleagues [75] 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 [32]. 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 [34]).

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, e.g., $\mathbf{E}S_A = s + b_A$ or $\mathbf{E}S_V = s + b_V$ (see [68] 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* [48, 87] 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, 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 auditory-visual 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 [103], x_V, x_A stand for the perceptual values of the event, i.e., 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' rule,

$$(7) \quad p(C | x_V, x_A) = \frac{p(x_V, x_A | C) p(C)}{p(x_V, x_A)}.$$

Going over to the ratio and taking logarithms, we obtain

$$(8) \quad \log \frac{\Pr(C = 1 | x_V, x_A)}{\Pr(C = 2 | x_V, x_A)} = \log \frac{p(x_V, x_A | C = 1)}{p(x_V, x_A | C = 2)} + \log \frac{\Pr(C = 1)}{\Pr(C = 2)}.$$

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

$$(9) \quad p(x_V, x_A | C = 1) = \int p(x_V, x_A | s) p(s) ds = \int p(x_V | s) p(x_A | s) p(s) ds.$$

For the denominator, marginalization is across both possible positions,

$$(10) \quad \begin{aligned} p(x_V, x_A | C = 2) &= \int \int p(x_V, x_A | s_V, s_A) p(s_V, s_A) ds_V ds_A \\ &= \left(\int p(x_V | s_V) p(s_V) ds_V \right) \left(\int p(x_A | s_A) p(s_A) ds_A \right). \end{aligned}$$

Explicit expressions for the likelihood functions under the above assumption of independent Gaussians are found in [48]. Under these assumptions, the optimal estimates for the common cause model are

$$(11) \quad \hat{s}_{A,C=1} = \hat{s}_{V,C=1} = \frac{\frac{x_A}{\sigma_A^2} + \frac{x_V}{\sigma_V^2} + \frac{\mu_p}{\sigma_p^2}}{\frac{1}{\sigma_A^2} + \frac{1}{\sigma_V^2} + \frac{1}{\sigma_p^2}};$$

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

$$(12) \quad \hat{s}_{A,C=2} = \frac{\frac{x_A}{\sigma_A^2} + \frac{\mu_p}{\sigma_p^2}}{\frac{1}{\sigma_A^2} + \frac{1}{\sigma_p^2}} \quad \text{and} \quad \hat{s}_{V,C=2} = \frac{\frac{x_V}{\sigma_V^2} + \frac{\mu_p}{\sigma_p^2}}{\frac{1}{\sigma_V^2} + \frac{1}{\sigma_p^2}}.$$

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, $p(C | x_V, x_A)$. 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 [48]:

$$(13) \quad \hat{s}_A = p(C = 1 | x_V, x_A) \hat{s}_{A,C=1} + [1 - p(C = 1 | x_V, x_A)] \hat{s}_{A,C=2}$$

$$(14) \quad \hat{s}_V = p(C = 1 | x_V, x_A) \hat{s}_{V,C=1} + [1 - p(C = 1 | x_V, x_A)] \hat{s}_{V,C=2}.$$

Note that these overall estimates are no longer linear combinations of x_V and x_A since 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:

$$(15) \quad p(\hat{s}_V | s_V, s_A) = \int \int p(\hat{s}_V | x_V, x_A) p(x_V | s_V) p(x_A | s_A) dx_V dx_A,$$

with an analogous expression for the auditory positions.

Two alternative decision strategies have been investigated in [103]. 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; e.g., for the auditory estimate,

$$(16) \quad \hat{s}_A = \begin{cases} \hat{s}_{A,C=1} & \text{if } p(C = 1 | x_V, x_A) > .5; \\ \hat{s}_{A,C=2} & \text{if } p(C = 1 | x_V, x_A) \leq .5. \end{cases}$$

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:

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

where u is from a standard uniform distribution;

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 Shams and collaborators [103], 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 [102].

Combining Bayesian causal inference with fMRI. In a landmark study by Rohe and Noppeney [81] 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 [46], 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 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. However, by assuming unbiasedness 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 [32] 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 [32]. Bias is assumed to be additive, that is, the biased sensory signals are

$$(18) \quad S_V = S_{W_V} + B_V \quad \text{and} \quad S_H = S_{W_H} + B_H,$$

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 $\mathcal{N}((S_V, S_H), \Sigma_z)$,

$$(19) \quad p(z_V, z_H | S_V, S_H) = \frac{1}{N_1} \exp \left[-\frac{1}{2} \left(\frac{(z_V - S_V)^2}{\sigma_V^2} + \frac{(z_H - S_H)^2}{\sigma_H^2} \right) \right]$$

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 Equation (18). The index MLE (maximum likelihood estimate) refers to the estimates as inferred directly from the sensory signals. The discrepancy between the signal estimates,

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

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 [32], 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' rule:

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

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 Figure 1), a special case of the upper Fréchet bound

([54], p. 155). This singular distribution can be approximated by

$$(20) \quad p(S_V, S_H) = \frac{1}{N_2} \exp \left\{ -\frac{1}{2} \frac{(S_V - S_H)^2}{\sigma_x^2} \right\}$$

with $\sigma_x \rightarrow 0$. Here, σ_x^2 measures the spread of the joint distribution function (middle row of Figure 1) and N_2 is another normalization constant. According to the Bayes' rule, the posterior distribution will also have probability mass only on the diagonal. Because of unbiasedness, the MAP estimate for the posterior distribution is shifted, relative to the MLE estimate of the likelihood function, in the direction of α (arrow in Figure 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 Equation (18) and $S_{W_V} = S_{W_H}$. The prior is the same as in Equation (20) 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 posterior estimate of the sensory signals, $(\hat{S}_V^{MAP}, \hat{S}_H^{MAP})$ as well as the best possible estimate of the discrepancy as a result of the bias $\hat{D}^{MAP} = \hat{S}_V^{MAP} - \hat{S}_H^{MAP} = \hat{B}_V - \hat{B}_H$. The posterior estimate is shifted with respect to the likelihood \hat{S}^{MLE} (see arrow in Figure 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 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, i.e. thus non-informative, and the *maximum a-posteriori* (MAP) estimate is given by the maximum likelihood estimate (MLE) of distribution (19) (bottom row of Figure 1),

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

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 $(\hat{B}_V^{MAP}, \hat{B}_H^{MAP})$ (for details we refer to [32]).

The predictions of this model, regarding both bias and variance, have been confirmed, e.g., by an experimental study of perceived numerosity of visual and haptic events [6]. Modifications of the model, relaxing some of the assumptions, e.g., allowing for correlated noise distributions of the perceptual estimates are presented in [32] 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 [18].

Some criticism of cue combination models. Rosas and Wichmann [82] 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.” ([82], p.147). This 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, e.g., to study under what circumstances behavior may actually result from simpler heuristic “non-optimal” cue combination rule.

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

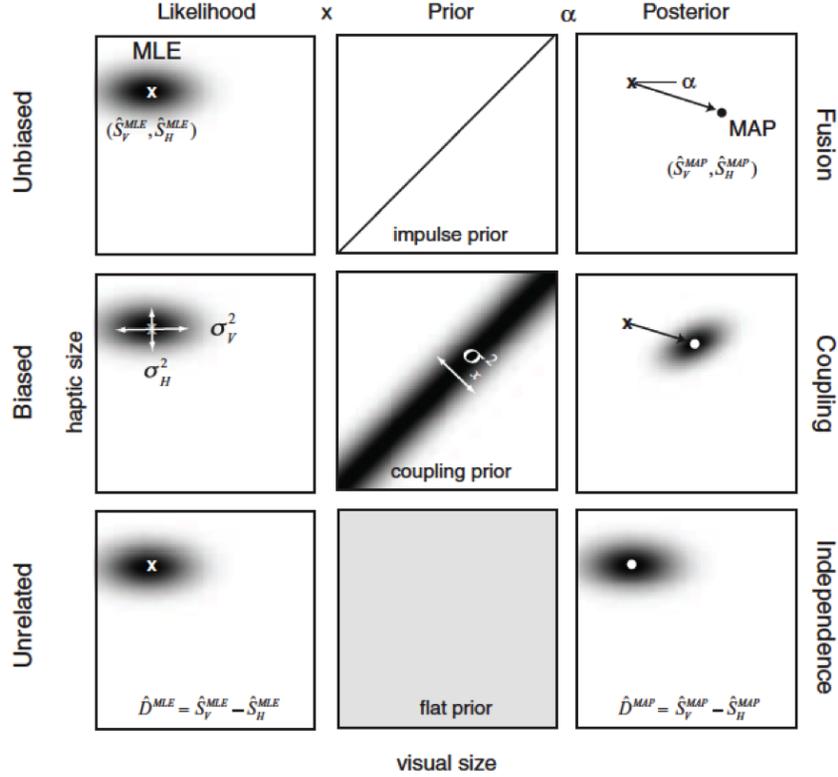


FIGURE 1. The combination of visual and haptic measurements with different prior distributions. (Left column) Likelihood functions resulting from noise with standard deviation σ_V twice as large as σ_H ; marks the maximum-likelihood estimate (MLE) of the sensory signals $(\hat{S}_V^{MLE}, \hat{S}_H^{MLE})$. (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 ($\sigma_x^2 \rightarrow \infty$). (Right column) Posterior distributions, which are the product of the likelihood and prior distributions according to Bayes rule. The dot marks the maximum a-posteriori (MAP) estimate $(\hat{S}_V^{MAP}, \hat{S}_H^{MAP})$. Black arrows correspond to the shift in the MAP estimate relative to the MLE estimate. 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 Ernst, 2012, p.531).

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 both in time and space. It has indeed been shown that that multisensory cue integration is statistically optimal (in the sense discussed above) only when signals are temporally correlated [73, 75]. The principle of correlation detection is well established in several areas of neuroscience, including vision [70] and hearing [43].

The multisensory correlation detector (MCD) model. An early model of multisensory synchrony detection relying on cross-correlation operations was proposed by Fujisaki and Nishida [37]. However, up to date the most comprehensive correlation detector model for multisensory integration has recently been developed by Parise and colleagues [72]. It is based on the Hassenstein-Reichardt model for local motion detection [42], commonly referred to as the Reichardt detector [99].

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 Figure 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, e.g. [36], and [4] 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

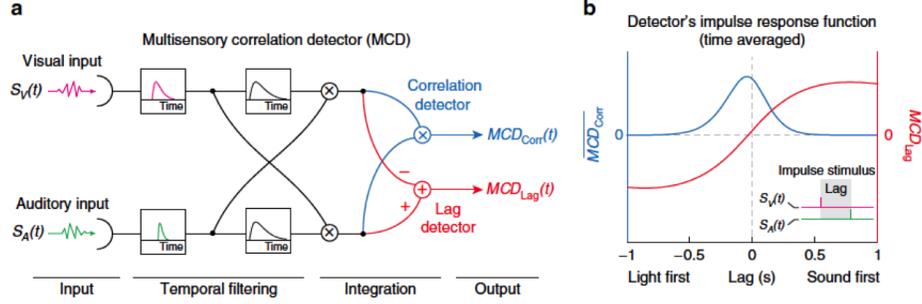


FIGURE 2. Schematic representation of the multisensory correlation detector (MCD) model (see text). From Parise & Ernst, 2016

and early unisensory processing). The filters are the same as in [7], i.e.,

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

where *mod* stands for the visual or auditory modality and τ_V, τ_A is the temporal constant of the filter. The next stage parallels the Reichardt detector in that these filtered signals are fed into two mirror-symmetric 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 Figure 2a). Specifically, for subunits (u_1, u_2) and signals $(S_V(t), S_A(t))$

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

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:

$$\begin{aligned} MCD_{Corr}(t) &= u_1(t) \times u_2(t) \\ MCD_{Lag}(t) &= -u_1(t) + u_2(t), \end{aligned}$$

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 Figure 2b).

Parise and Ernst [72] 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 was analyzed using psychophysical *reverse correlation* techniques. These analyses were performed on the cross-correlation profile of the signals since 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 (that is, light vs. sound lead). In order 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 [72]). 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 [75], make a clear case that, in a noisy environment, the brain tends to infer causality (e.g., unity of events) from correlation. However, 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, i.e., 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 $F_V(t)$. 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., [24]).

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 [79]. For simplicity, we here equate observable RTs with the finishing times of the “race”, neglecting any additional components of RT like motor processing, etc.. Assuming random variability of the finishing times, mean RT in the crossmodal condition with $\text{SOA} = \tau$, 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:

$$(21) \quad \mathbf{E}RT_{VA+\tau} \equiv \mathbf{E} \min\{RT_V, RT_{A+\tau}\} \leq \min\{\mathbf{E}RT_V, \mathbf{E}RT_{A+\tau}\}.$$

The equivalence on the left-hand side of Equation (21) expresses assumption (ii) of the race model, and the inequality follows as a special case of Jensen’s inequality (e.g., [3]). 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 [49], pp. 129–130): Note that each context \mathcal{V} , \mathcal{A} , or \mathcal{VA} refers to a different sample space (and corresponding σ -algebra), since 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,

$$\begin{aligned} H_{VA}(s, t) &\equiv \Pr(RT_V \leq s, RT_A \leq t) \\ &= F_V(s) F_A(t), \end{aligned}$$

for all nonnegative s, t . In particular,

$$(22) \quad H_{VA}(s, \infty) = F_V(s) \text{ and } H_{VA}(\infty, t) = F_A(t).$$

For the stochastically independent race model,

$$(23) \quad \begin{aligned} F_{VA}(t) &= \Pr(\min\{RT_V, RT_A\} \leq t) \\ &= \Pr(RT_V \leq t) + \Pr(RT_A \leq t) - \Pr(RT_V \leq t) \Pr(RT_A \leq t). \end{aligned}$$

Observing crossmodal RT distributions exceeding the right-hand side of Equation (23) at any time point t leads to a rejection of the independent race model. Note, however, that the equalities in (22) 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* [11, 96], is not empirically testable because $H_{VA}(s, t)$ is not observable in paradigm RSP. As recently demonstrated in [62], context invariance is

an essential part of the race model concept and dropping it makes the race model unfalsifiable.

The race model inequality. Rewriting Equation (23) without the independence assumption,

$$F_{VA}(t) = \Pr(RT_V \leq t) + \Pr(RT_A \leq t) - H_{VA}(t, t). \quad (24)$$

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

$$F_{VA}(t) \leq \Pr(RT_V \leq t) + \Pr(RT_A \leq t) \\ = F_V(t) + F_A(t). \quad (25)$$

This is the *race model inequality*. As observed by J. Miller [60], it constitutes a stronger test of the model than a test based on the independent race model (Equation 23) or on the level of means using Inequality (21). 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 [11, 13]. Generalizations of the race model inequality to three different modalities have been discussed in [20, 17].

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 and Minakata [39] discuss a variety of statistical methods in wide use. Many of these tests are somewhat questionable being based on multiple testing, 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 [15]. It turns out this test simply compares $\mathbf{E}RT_{VA}$ with the expected value of random variable $\min\{V, A\}$ under maximal negative dependence, that is, SM^- in Equation (3).

Violation of Equation (25) 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 [38]. In a recent study employing a large set of natural images and sounds, Juan and colleagues found a large range of violations and non-violations of the race model inequality as a function of the intrinsic physical parameters of the stimuli in both humans and monkeys [45].

Coactivation models for reaction time. Coactivation models for reaction time have been defined by Jeff Miller [60] 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, i.e., 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 assumed as well.

If one were to assume 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 c th count to occur note the well known identity

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

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 c th count then simply is $\mathbf{E}S_c = c/\lambda$. Here, c is often interpreted as a bias parameter describing the subjects 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 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 [21]. Fits of the Poisson counter model to crossmodal RTs have generally been quite satisfying, at least at the level of RT means [85, 19, 23]. If an additive (motor) component is added to the decision RT, variances can also be fit satisfactorily [86]. 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 version of the coactivation idea.

Continuous accumulation: Diffusion models. In the diffusion model for crossmodal RTs, the 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., [86]). For non-overlapping 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, *first passage time* T , is

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

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

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

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

$$\mathbf{E}_{VA} = c/(\mu_V + \mu_A),$$

Although the two models are fundamentally different from a theoretical viewpoint, predictions of the diffusion model and the Poisson superposition models are clearly indistinguishable as long as only expected decision times are considered. Thus, the diffusion model was developed further in two different directions. Schwarz [86] derived expressions for the first passage time as function of stimulus onset asynchrony (SOA) and added a non-independent motor component to obtain an excellent fit to the classic data set of Miller [61]. Using a matrix approximation approach to diffusion processes, Diederich [21] 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, i.e., 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, since 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 ([80, 97]). As before in fitting the Poisson counter model, Diederich 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 ([25])

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

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 because errors rates are typically kept very low in these RT experiments. On the other hand, in paradigms other than multisensory integration, both diffusion models and Poisson superposition models have been extended to predict error rates and error RTs (e.g., [22, 100, 89]). Another more elaborated version of drift was proposed in a recent diffusion model by Drugowitsch and colleagues

for a visual/vestibular heading discrimination task[31]. 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

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

where $k_{vis}(c)$ 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 [31]). 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's 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 our 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 [14, 25] 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 sub-processes 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, nonnegative 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

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

where the presentation of the visual stimulus is arbitrarily defined as the 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 [28, 29, 26, 27], 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 (mixture of) ex-Gaussian distributions, which are quite common in RT modeling (**REF**).

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 , $\mathbf{E}[RT_{V\tau A}]$, is computed conditioning on event I (integration)

occurring or not,

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

Here, I^c denotes the complementary event to I , μ is short for $\mathbf{E}[S_2|I]$, and Δ stands for $\mathbf{E}[S_2|I^c] - \mathbf{E}[S_2|I]$. 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 $\mathbf{E}[RT_{V\tau A}]$ 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,

$$\mathbf{E}[RT_V] = \mathbf{E}[V] + \mathbf{E}[S_2|I^c] \quad \text{and} \quad \mathbf{E}[RT_A] = \mathbf{E}[A] + \mathbf{E}[S_2|I^c].$$

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 [79]:

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

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

$$\min\{\mathbf{E}[RT_V], \mathbf{E}[RT_A] + \tau\} - \mathbf{E}[RT_{V\tau A}] = SFE + P(I) \times \Delta.$$

Thus, crossmodal *facilitation* observed in a redundant signals task may be due either to multisensory integration or statistical facilitation, or 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, e.g., temporal order judgment (TOJ) or simultaneity (synchronicity) judgment ([101] for a recent review). As recently shown [30], 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 reanalysis ([30]) of data from Mégevand and colleagues [56] 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 [41]. 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 audio-visual discrepancy after a single trial presentation [104, 57]. Clearly, further research on the mechanisms underlying both fast and slow recalibration is called for, and the TWIN modeling framework also needs to be extended in order to capture dependencies across trials.

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 to behavioral performance, in particular cue integration (for a comprehensive recent review, see [34]). 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 [50] 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

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

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:

$$(27) \quad p(s|\mathbf{r}) \propto p(\mathbf{r}|s)p(s),$$

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 (27) holds even if the prior $p(s)$ is flat (uniform). As formulated in [51], 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. On the other hand, neurons do not exactly follow Poisson and are correlated. More general approaches have been proposed, in particular so called *Poisson-like* variability, i.e., the exponential family with linear sufficient statistics [50].

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

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

Mean spike count is assumed to be $\lambda_i = g f_i(s)$ 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 (as function of s) $p(\mathbf{r}|s)$ equals

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

From (27), 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 since 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,

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

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, i.e. 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:

$$(28) \quad \mathbf{r}_{AV} = \mathbf{r}_A + \mathbf{r}_V.$$

The output pattern \mathbf{r}_{AV} will still obey independent Poisson variability across many trials, since the sum of two Poisson processes is again Poisson. The mean activity of the i th 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

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

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 [50]). The PPC framework was used in analyzing recordings from area MSTd in monkeys in a visual-vestibular heading task [35]. 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 Equation (28) and the minimum variance linear combination rule of the standard Gaussian cue combination model of Equation (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 [52]. Nevertheless, Fetsch and colleagues [34] argue convincingly that three distinct uses of the term “weights” need to be distinguished: 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 is measured in most single-unit studies of multisensory integration and, finally, the relationship between those neural weights and the perceptual weights measured in psychophysical tasks is mostly unknown yet. In this context, Ma and Rahmati [53] presented a PPC implementation of the Bayesian causal inference model but concluded that the resulting architecture is unrealistic since 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 [10]). 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):

$$(29) \quad R = \frac{E^n}{\alpha^n + \left(\frac{1}{N}\right) \sum_{j=1}^N E_j^n}.$$

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 and colleagues [69] developed a 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,

$$(30) \quad E = d_1 I_1(x_0, y_0) + d_2 I_2(x_0, y_0).$$

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 Equation (29). 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 [1, 84]. They go beyond models like Ohshiro et al. [69] (i) in representing the temporal profile of multisensory enhancement (or suppression) [83], (ii) in explicitly taking into account the finding that multisensory integration in the SC is mediated by descending inputs from association cortex [44], and in addressing questions about the emergence and maturation of multisensory integration in dependence of the environment [94]. In a recent computational model, the continuous-time

multisensory model (CTMM) [64], 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 [72] 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. The basic hypothetical assumption of the model developed in [8] 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 cross-modal and within-modal stimuli (for details, see [8]). A recent model by the same authors [98] 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.

APPENDIX (PROOF OF PROPOSITION 1)

The proof uses a special case of the Cauchy-Schwarz inequality (see e.g. [65], p.30) that follows from simple algebraic transformations:

Lemma 2. 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,

$$(31) \quad (w_V + w_H)^2 \leq (w_V^2 \sigma_V^2 + w_H^2 \sigma_H^2) \left(\frac{1}{\sigma_V^2} + \frac{1}{\sigma_H^2} \right)$$

The first term on the right-hand side is the variance to be minimized and the other terms do not depend on choice of the weights. A minimum is obviously

achieved when Inequality 31 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.

REFERENCES

1. J.C. Alvarado, B.A. Rowland, T.R. Stanford, and B.E. Stein, *A neural network model of multisensory integration also accounts for unisensory integration in superior colliculus*, *Brain Research* **1242** (2008), 13–23.
2. M. S. Beauchamp, *Statistical criteria in fMRI studies of multisensory integration*, *Neuroinformatics* **3** (2005), 93–113.
3. R. N. Bhattacharya and E.C. Waymire, *Stochastic processes with applications*, Wiley, New York, 1990.
4. A. Borst and T. Euler, *Seeing things in motion: models, circuits, and mechanisms*, *Neuron* **71** (2011), 974–994.
5. A.J. Bremner, D.J. Lewkowicz, and C. Spence (eds.), *Multisensory development*, Oxford University Press, Oxford, UK, 2012.
6. J.-P. Bresciani, F. Dammeier, and M.O. Ernst, *Vision and touch are automatically integrated for the perception of sequences of events*, *Journal of Vision* **6** (2006), 554–564.
7. D. Burr, O. Silva, G.M. Cichini, M. S. Banks, and M.C. Morrone, *Temporal mechanisms of multimodal binding*, *Proceedings of the Royal Society B* **276** (2009), 1761–1769, doi: 10.1098/rspb.2008.1899.
8. Cuppini C, M. Ursino, E. Magosso, B.A Rowland, and B.E. Stein, *An emergent model of multisensory integration in superior colliculus neurons*, *Frontiers in Integrative Neuroscience* (2010), doi: 10.3389/fnint.2010.00006.
9. G. Calvert, C. Spence, and B.E. Stein (eds.), *Handbook of multisensory processes: Modeling the time course of multisensory perception in manual and saccadic responses*, MIT Press, Cambridge, MA, 2004.
10. M. Carandini and D.J. Heeger, *Normalization as a canonical neural computation*, *Nature Reviews Neuroscience* **13** (2012), 51–62.
11. H. Colonius, *Possibly dependent probability summation of reaction time*, *Journal of Mathematical Psychology* **34** (1990), no. 3, 253–275.
12. ———, *Behavioral measures of multisensory integration: Bounds on bimodal detection probability*, *Brain Topography* **28** (2015), no. 1, 1–4.

13. ———, *An invitation to coupling and copulas, with applications to multisensory modeling*, *Journal of Mathematical Psychology*, dx.doi.org/10.1016/j.jmp.2016.02.004 **74** (2016), 2–10.
14. H. Colonius and A. Diederich, *Multisensory interaction in saccadic reaction time: a time-window-of-integration model*, *Journal of Cognitive Neuroscience* **16** (2004), 1000–1009.
15. ———, *The race model inequality: interpreting a geometric measure of the amount of violation.*, *Psychological Review* **113** (2006), no. 1, 148–154.
16. ———, *Measuring multisensory integration: from reaction times to spike counts*, *Scientific Reports* **7** (2017), no. 1, 3023, <http://dx.doi.org/10.1038/s41598-017-03219-5>.
17. H. Colonius, F.H. Wolff, and A. Diederich, *Trimodal race model inequalities in multisensory integration: I. Basics*, *Frontiers in Psychology* (2017), 10.3389, doi: 10.3389/fpsyg.2017.01141.
18. N.B. Debats, M.O. Ernst, and H. Heuer, *Perceptual attraction in tool use: evidence for a reliability-based weighting mechanism*, *Journal of Neurophysiology* **117** (2017), 1569–1580.
19. A. Diederich, *Intersensory facilitation: Race, superposition, and diffusion models for reaction time to multiple stimuli*, Peter Lang, Frankfurt, 1992.
20. ———, *Probability inequalities for testing separate activation models of divided attention*, *Perception & Psychophysics* **52** (1992), no. 6, 714–716.
21. ———, *Intersensory facilitation of reaction time: Evaluation of counter and diffusion coactivation models.*, *Journal of Mathematical Psychology* **39** (1995), 197–215.
22. ———, *MDFT account of decision making under time pressure*, *Psychonomic Bulletin & Review* **10** (2003), no. 1, 157–166.
23. A. Diederich and H. Colonius, *A further test of the superposition model for the redundant signals effect in bimodal detection*, *Perception & Psychophysics* **50** (1991), 83–83.
24. ———, *Bimodal and trimodal multisensory enhancement: effects of stimulus onset and intensity on reaction time*, *Perception and Psychophysics* **66** (2004), 1388–1404.
25. ———, *Modeling the time course of multisensory interaction in manual and saccadic responses*, *Handbook of multisensory processes* (G. Calvert, C. Spence, and B.E. Stein, eds.), MIT Press, Cambridge, MA, 2004, pp. 395–408.
26. ———, *Modeling spatial effects in visual-tactile reaction time*, *Perception and Psychophysics* **69** (2007), no. 1, 56–67.
27. ———, *Why two "distractors" are better than one: Modeling the effect on non-target auditory and tactile stimuli on visual saccadic reaction time*, *Experimental Brain Research* **179** (2007), 43–54.
28. ———, *Crossmodal interaction in saccadic reaction time: Separating multisensory from warning effects in the time window of integration model*, *Experimental Brain Research* **186** (2008), no. 1, 1–22.
29. ———, *When a high-intensity "distractor" is better than a low-intensity one: Modeling the effect of an auditory or tactile nontarget stimulus on visual saccadic reaction time*, *Brain Research* **1242** (2008), 219–230.

30. ———, *The time window of multisensory integration: Relating reaction times and judgments of temporal order*, *Psychological Rev* **122** (2015), no. 2, 232–241.
31. J. Drugowitsch, G.C. DeAngelis, E. M. Klier, D.E. Angelaki, and A. Pouget, *Optimal multisensory decision-making in a reaction-time task*, *elife* (2014), e03005, doi: 10.7554/eLife.03005.
32. M.O. Ernst, *Optimal multisensory integration: assumptions and limits.*, *The New Handbook of Multisensory Processes* (B.E. Stein, ed.), MIT Press, Cambridge, MA, 2012, pp. 527–544.
33. M.O. Ernst and M. S. Banks, *Humans integrate visual and haptic information in a statistically optimal fashion*, *Nature* **415** (2002), 429–433.
34. C.R. Fetsch, G.C. DeAngelis, and D.E. Angelaki, *Bridging the gap between theories of sensory cue integration and the physiology of multisensory neurons*, *Nature Reviews Neuroscience* **14** (2013), 429–442.
35. C.R. Fetsch, A. Pouget, G.C. DeAngelis, and D.E. Angelaki, *Neural correlates of reliability-based cue weighting during multisensory integration*, *Nature Neuroscience* **15** (2011), 146–154.
36. Y.E. Fisher, M. Sillies, and T.R. Clandinin, *Orientation selectivity sharpens motion detection in *Drosophila**, *Neuron* **88** (2015), 390–402.
37. W. Fujisaki and S. Nishida, *Feature-based processing of audio-visual synchrony perception revealed by random pulse trains*, *Vision Research* **47** (2007), 1075–1093.
38. M. Gondan, K. Lange, F. Rösler, and B. Röder, *The redundant target effect is affected by modality switch costs*, *Psychonomic Bulletin & Review* **11** (2004), no. 2, 307–313.
39. M. Gondan and K. Minakata, *A tutorial on testing the race model inequality*, *Attention, Perception, and Psychophysics* **78** (2016), no. 3, 723–735, doi: 10.3758/s13414-015-1018-y.
40. D.M. Green and J.A. Swets, *Signal detection theory and psychophysics*, Wiley, New York, 1966.
41. V. Harrar, L. Harris, and C. Spence, *Multisensory integration is independent of perceived simultaneity*, *Experimental Brain Research* (2016).
42. V. Hassenstein and W. Reichardt, *System theoretical analysis of time, sequence and sign analysis of the motion perception of the snout-beetle *chlorophanus**, *Zeitschrift für Naturforschung* **11** (1956), 513–524, (in German).
43. L.A. Jeffress, *A place theory of sound localization*, *Journal of Comparative and Physiological Psychology* **41** (1948), no. 1, 35–39.
44. W. Jiang, M.T. Wallace, H. Jiang, J.W. Vaughan, and B.E. Stein, *Two cortical areas mediate multisensory integration in superior colliculus neurons*, *Journal of Neurophysiology* **85** (2001), 506–522.
45. C. Juan, C. Cappe, B. Alric, B. Roby, S. Gilardeau, P. Barone, and P. Girard, *The variability of multisensory processes of natural stimuli in human and non-human primates in a detection task*, *PLoS ONE* **12** (2017), no. 2, e0172480, doi:10.1371/journal.pone.0172480.
46. C. Kayser and L. Shams, *Multisensory causal inference in the brain*, *PLoS Biology* **13** (2015), no. 2, e1002075, doi:10.1371/journal.pbio.1002075.

47. D.C. Knill and W. Richards (eds.), *Perception as Bayesian inference*, Cambridge University Press, Cambridge, UK, 1996.
48. K.P. Körding, U.R. Beierholm, W.J. Ma, S. Quartz, J.B. Tenenbaum, and L. Shams, *Causal inference in multisensory perception*, PLoS ONE **2** (2007), no. 9, e943. doi:10.1371/journal.pone.0000943.
49. R.D. Luce, *Response times: Their role in inferring elementary mental organization*, Oxford University Press, New York, NY, 1986.
50. W.J. Ma, J.M. Beck, P.E. Latham, and A. Pouget, *Bayesian inference with probabilistic population codes*, Nature Neuroscience **9** (2006), 1432–1438.
51. W.J. Ma, J.M. Beck, and A. Pouget, *Spiking networks for Bayesian inference and choice*, Current Opinion in Neurobiology **18** (2008), 217–222.
52. W.J. Ma and A. Pouget, *Linking neurons to behavior in multisensory perception: A computational review*, Brain Research (2008), no. 1242, 4–12.
53. W.J. Ma and M. Rahmati, *Towards a neural implementation of causal inference in cue combination*, Multisensory Research **26** (2013), 159–176.
54. D.D. Mari and S. Kotz, *Correlation and dependence*, Imperial College Press, London, 2001.
55. H McGurk and J MacDonald, *Hearing lips and seeing voices*, Nature **264** (1976), no. 5588, 746–8.
56. P. Mégevand, S. Molholm, A. Nayak, and J. J. Foxe, *Recalibration of the multisensory temporal window of integration results from changing task demands*, PLoS ONE **8** (2013), no. 8, e71608.
57. C. Mendonca, A. Escher, S. van de Par, and H. Colonius, *Predicting auditory space calibration from recent multisensory experience*, Experimental Brain Research **233** (2015), 1983–1991.
58. M.A. Meredith, *On the neuronal basis for multisensory convergence: a brief overview*, Cognitive Brain Research **14** (2002), 31–40.
59. M.A. Meredith and B.E. Stein, *Interactions among converging sensory inputs in the superior colliculus*, Science **221** (1983), 389–391.
60. J. Miller, *Divided attention: Evidence for coactivation with redundant signals.*, Cognitive Psychology **14** (1982), 247–279.
61. ———, *Timecourse of coactivation in bimodal divided attention*, Perception & Psychophysics **40** (1986), 331–343.
62. ———, *Statistical facilitation and the redundant signals effect: What are race and coactivation models?*, Attention, Perception, and Psychophysics **78** (2016), 516–519.
63. R.L. Miller, S.R. Pluta, B.E. Stein, and B.A. Rowland, *Relative unisensory strength and timing predict their multisensory product*, The Journal of Neuroscience **35** (2015), no. 13, 5213–5220.
64. R.L. Miller, B.E. Stein, and B.A. Rowland, *Multisensory integration uses a real time unisensory-multisensory transform*, Journal of Neuroscience (2017), DOI: 10.1523/JNEUROSCI.2767-16.2017.
65. D.S. Mitrinović, *Analytic inequalities*, Die Grundlehren der mathematischen Wissenschaften, vol. 165, Springer-Verlag, Berlin, 1970.

66. M.M. Murray and M.T. Wallace (eds.), *The neural bases of multisensory processes*, Frontiers in Neuroscience, CRC Press, Boca Rato, FL, 2012.
67. M.J. Naumer and J. Kayser (eds.), *Multisensory object perception in the primate brain*, Springer-Verlag, New York, NY, 2010.
68. B. Odegaard, D.R. Wozny, and L. Shams, *Biases in visual, auditory, and audiovisual perception of space*, PLoS Computational Biology **11** (2015), no. 12, e1004649. doi:10.1371/journal.pcbi.1004649.
69. T. Ohshiro, D.E. Angelaki, and G.C. DeAngelis, *A normalization model of multisensory integration*, Nature Neuroscience **14** (2011), no. 6.
70. I. Ohzawa, *Mechanisms of stereoscopic vision: the disparity energy model*, Current Opinion in Neurobiology **8** (1998), 509–515.
71. İ. Oruç, L.T. Maloney, and M. S. Landy, *Weighted linear cue combination with possibly correlated error*, Vision **43** (2003), 2451–2468.
72. C.V. Parise and M.O. Ernst, *Correlation detection as a general mechanism for multisensory integration*, Nature Communications **7** (2016), no. 11543, doi: 10.1038/ncomms11543.
73. C.V. Parise, V. Harrar, M.O. Ernst, and C. Spence, *Cross-correlation between auditory and visual signals promotes multisensory integration*, Multisensory Research **26** (2013), 307–316.
74. C.V. Parise and C. Spence, *'When birds of a feather flock together': Synesthetic correspondences modulate audiovisual integration in non-synesthetes*, PloS One **4**(5) (2009), no. e5664, doi:10.1371/journal.pone.0005664.
75. C.V. Parise, C. Spence, and M.O. Ernst, *When correlation implies causation in multisensory integration*, Current Biology **22** (2012), 1–4.
76. T.J. Perrault Jr, J.W. Vaughan, B.E. Stein, and M.T. Wallace, *Superior colliculus neurons use distinct operational modes in the integration of multisensory stimuli*, Journal of Neurophysiology **93** (2005), 2575–2586.
77. L.C. Populin and T.C. Yin, *Bimodal interactions in the superior colliculus of the behaving cat*, Journal of Neuroscience **22** (2002), 2826–2834.
78. A. Pouget, S. Deneve, and J.R. Duhamel, *A computational perspective on the neural basis of multisensory spatial representations*, Nature Reviews Neuroscience **3** (2002), 741–747.
79. D.H. Raab, *Statistical facilitation of simple reaction time*, Transactions of the New York Academy of Sciences **24** (1962), 574–590.
80. L.M. Ricciardi, *Diffusion processes and related topics in biology*, Springer Verlag, Berlin, 1977.
81. T. Rohe and U. Noppeney, *Cortical hierarchies perform bayesian causal inference in multisensory perception*, PLoS Biology **13** (2015), no. 2, e1002073, doi:10.1371/journal.pbio.1002073.
82. P. Rosas and F.A. Wichmann, *Cue combination: Beyond optimality*, Sensory cue integration (J. Trommershäuser, K. Körding, and M.S. Landy, eds.), Oxford University Press, 2011, pp. 144–152.

83. B.A. Rowland, S. Quessy, T.R. Stanford, and B.E. Stein, *Multisensory integration shortens physiological response latencies*, *Journal of Neuroscience* **27** (2007), 5879–5884.
84. B.A. Rowland and B.E. Stein, *A model of the temporal dynamics of multisensory enhancement*, *Neuroscience and Biobehavioral Reviews* **41** (2014), 78–84.
85. W. Schwarz, *A new model to explain the redundant-signals effect*, *Perception & Psychophysics* **46** (1989), no. 5, 490–500.
86. ———, *Diffusion, superposition, and the redundant-targets effect*, *Journal of Mathematical Psychology* **38** (1994), 504–520.
87. L. Shams and U.R. Beierholm, *Causal inference in perception*, *Trends in Cognitive Sciences* **14** (2010), 425–432.
88. M.L. Shaw, *Attending to multiple sources of information: I: The integration of information in decision making*, *Cognitive Psychology* **14** (1982), no. 3, 353–409.
89. P.L. Smith and T. van Zandt, *Time-dependent poisson counter models of response latency in simple judgment*, *British Journal of Mathematical and Statistical Psychology* **53** (2000), 293–315.
90. B.E. Stein (ed.), *The new handbook of multisensory processes*, MIT Press, Cambridge, MA, 2012.
91. B.E. Stein, D. Burr, C. Constantinidis, P.J. Laurienti, M. Meredith, T. Perrault Jr, R. Ramachandran, B. Roeder, B. Rowland, K. Sathian, C.E. Schroeder, L. Shams, T. Stanford, M. Wallace, L. Y, and D.J. Lewkowicz, *Semantic confusion regarding the development of multisensory integration: a practical solution*, *European Journal of Neuroscience* **31** (2010), 1713–1720.
92. B.E. Stein and M.A. Meredith, *The merging of the senses*, MIT Press, 1993.
93. B.E. Stein, T.R. Stanford, R. Ramachandran, T.J. Perrault Jr, and B.A. Rowland, *Challenges in quantifying multisensory integration: alternative criteria, models, and inverse effectiveness*, *Experimental Brain Research* **198** (2009), 113–126.
94. B.E. Stein, T.R. Stanford, and B.A. Rowland, *Development of multisensory integration from the perspective of the individual neuron*, *Nature Reviews Neuroscience* **15** (2014), 520–535.
95. J.W. Todd, *Reaction to multiple stimuli*, *Archives of Psychology* No. 25. Columbia Contributions to Philosophy and Psychology **XXI** (1912), no. 8, New York: The Science Press.
96. J.T. Townsend and G. Nozawa, *Spatio-temporal properties of elementary perception: An investigation of parallel, serial, and coactive theories*, *Journal of Mathematical Psychology* **39** (1995), 321–359.
97. H.C. Tuckwell, *Elementary applications of probability theory - with an introduction to stochastic differential equations*, 2nd ed., Chapman and Hall, London, 1995.
98. M. Ursino, C. Cuppini, and Mag, *Multisensory bayesian inference depends on synapse maturation during training: Theoretical analysis and neural modeling implementation*, *Neural Computation* **29** (2017), 735–782.
99. J.P.H. van Santen and G. Sperling, *Elaborated Reichardt detectors*, *Journal of the Optical Society of America A* **2** (1985), 300–321.

100. T. van Zandt, H. Colonius, and R.W. Proctor, *A comparison of two response time models applied to perceptual matching*, *Psychonomic Bulletin & Review* **7** (2000), no. 2, 208–256.
101. J. Vroomen and M. Keetels, *Perception of intersensory synchrony: A tutorial review*, *Attention, Perception, and Psychophysics* **72** (2010), no. 4, 871–884.
102. N. Vulkan, *An economist's perspective on probability matching*, *Journal of Economic Surveys* **14** (2000), no. 1, 101–118.
103. D.R. Wozny, U.R. Beierholm, and L. Shams, *Probability matching as a computational strategy used in perception*, *PLoS Computational Biology* **6** (2010), no. 8, e1000871. doi:10.1371/journal.pcbi.1000871.
104. D.R. Wozny and L. Shams, *Recalibration of auditory space following milliseconds of cross-modal discrepancy*, *Journal of Neuroscience* **31** (2011), no. 12, 4607–4612.