Measuring Multisensory Integration in Selected Paradigms

Adele Diederich Hans Colonius

This is a chapter for volume 3 of the New Handbook of Mathematical Psychology, edited by F.G. Ashby, H. Colonius, and E.N. Dzhafarov (Cambridge UP, 2023)

Contents

0.1 Overview

The investigation of processes involved in merging information from different sensory modalities has become the subject of research in many areas, including anatomy, physiology, and behavioral sciences. This field of research termed "multisensory integration" (MI) is flourishing, crossing borders between psychology and neuroscience. The focus of this chapter is on *measures* of multisensory integration based on numerical data collected from single neurons and in behavioral paradigms: spike numbers, reaction time, frequency of correct or incorrect responses in detection, recognition, and discrimination tasks. Defining that somewhat fuzzy term, it has been observed that at least some kind of *numerical* measurement assessing the strength of crossmodal effects is required. On the empirical side, these measures typically serve to quantify effects of various covariates on MI, like age, certain disorders (e.g., dyslexia), developmental conditions, training and rehabilitation, in addition to attention and learning. On the theoretical side, these measures often help to probe hypotheses about underlying integration mechanisms like optimality in combining information or inverse effectiveness, without necessarily subscribing to a specific model.

Given the important role of its neurophysiological basis, we start with a presentation of the major rules of integration observed in neural responses in the form of spike numbers elicited, and introduce numerical measures based on them. The essential role of the concept of "probability summation" in deriving measures satisfying certain "optimality" criteria emerges soon, and it reappears in later sections on measures based on response speed in different behavioral paradigms.[1]

Subsequently, measures based on accuracy are discussed in the context of signal detection theory, followed by measures developed within the broad area of audiovisual speech identification. A proposal for measuring integration efficiency based on the *Fechnerian Scaling* approach closes that section.

The number of models trying to reveal the mechanisms underlying MI at different levels of description, from the neural to the behavioral, is large and growing. In the corresponding section, we had to be very selective, and we primarily sketch models that help motivating a specific measure of integration.

In order to keep the presentation focused, measures suggested for multisensory "illusions", like the McGurk effect or the sound-induced flash illusion (typically, percentages), are not considered at all, nor are those derived from functional magnetic resonance imaging data sets. A list of all measures dis-

[1] Optimality is always defined here only in relation to a specific paradigm.

cussed in the chapter is found in the discussion section. Finally, the reader should not expect a balanced presentation of the large field of measuring multisensory integration; instead, we mainly consider those more or less related to our own work.

0.2 Measures of multisensory integration: introduction

0.2.1 Defining multisensory integration

Progress in MI is documented in several recent handbooks (see Section 0.8 for an overview of literature). Due to the large range of contexts – from neurophysiology to applied psychology and marketing, from single cells to food tastes – the field has been labeled in different ways, e.g. as "intersensory facilitation/enhancement", "intersensory/crossmodal interaction", or "multisensory integration", creating some semantic confusion among many researchers. In 2010, a group of authors, together with Barry Stein, one of the founders of the field in neuroscience, agreed upon defining "multisensory integration" as

> "*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." (Stein et al., 2010, p.1719).*

This broad definition does not commit to a specific model or experimental paradigm, nor to a criterion of optimality. Nevertheless, it requires some type of measure to assess whether the multisensory response is "significantly different" from the unisensory responses. Investigating such measures, as well as some models related to them, is the focus of this chapter.[2] Moreover, while the definition encompasses both facilitation and inhibition of the multisensory response, most measures presented here are formulated for the case of facilitation only and would need to be adapted to comprise inhibition.

0.2.2 Measuring multisensory integration

First, we introduce some needed notation. Beginning with the stimulus side, stimuli of a specific modality are labeled by s_A, s_V, s_T, for auditory, visual, and tactile (or somatosensory) stimuli, respectively, where further stimulus-specific information, like intensity, have to be added as needed. When a

[2] Note, however, that issues of testing *statistical* significance are not central to this chapter.

Acronym	Meaning
CRE	crossmodal response enhancement
E	expected value (mean)
FS	Fechnerian Scaling
FLMP	fuzzy logical model of perception
IE	integration efficiency
MI	multisensory integration
OUP	Ornstein-Uhlenbeck process
PRE	prelabeling (model)
PS	probability summation
RMI	race model inequality
RT	reaction time
SC	superior colliculus
SDT	signal detection theory
SFE	statistical facilitation effect
SOA	stimulus onset asynchrony
SRT	saccadic reaction time
TOJ	temporal order judgment
TWIN	time window of integration (model)
UI	unisensory balance
VE/AE	visual/auditory enhancement

Table 0.1 *Abbreviations used in the chapter*

label is only used as index of modality, we often omit the s part. A basic distinction to keep in mind is between a unisensory context where stimuli of a single modality, s_A, s_V, s_T, are presented, and a cross-sensory context where stimuli from two or more modalities are presented in a to-be-specified spatio-temporal arrangement. For concreteness, we refer to A, V, T as the unisensory context where only auditory, visual, or tactile stimuli are presented, respectively. Similarly, VA denotes a bisensory (visual-auditory) context with stimulus combinations labeled s_{VA} being presented, VAT a trisensory context with combined stimuli s_{VAT}, etc., where again further information about the specific presentation mode may have to be added. When the number of sensory modalities is not specified, we also use the label *crossmodal* (for context, condition, stimulus, response, etc.). Moreover, in this chapter mainly measures combining the visual and auditory modalities will be considered, but most of these would also apply to other modality combinations with minor modification.

Each time a specific auditory stimulus s_A, say, is presented, it will give rise to a unisensory response, e.g., a reaction time or a number of spikes within a certain time interval. Typically, these responses are considered as instantiation (realization) of some random variable, e.g. RT_A or N_A, respectively.

Similarly, a combination stimulus s_{VA} elicits bisensory responses considered as realizations of some random variables, RT_{VA} or N_{VA}. To simplify the exposition, we will neglect all experimental details for now.

At the sample level, a descriptive measure of MI has to relate the set of multisensory responses to the sets of unisensory responses; for example, how much differs the average auditory-visual response to the average auditory and average visual response? At the level of random variables, the MI measure should assess how, or how much, the distribution of responses to bisensory stimuli differs from the distributions to unisensory stimuli.

We define measures only at the level of probability distributions, the corresponding sample level measures are then easily derivable. In order to reduce the number of possible formats, one should consider necessary or desirable features of such a measure, denoted by CRE (crossmodal response enhancement/inhibition). We first state a few elementary properties any CRE measure of MI should have. The following list seems uncontroversial:

(i) (*Real-valued function*) CRE is a real-valued function of the crossmodal and unisensory empirical distributions, or of some parameter of these distributions (e.g. the mean);
(ii) (*No-integration case*) If the crossmodal distribution does not differ from one of the unisensory distributions, CRE equals zero;
(iii) (*facilitation-inhibition*) Negative values of CRE indicate crossmodal inhibition, positive values crossmodal facilitation.

Clearly, these features do not impose strong restrictions on the form of the measure; this does not come as a surprise, however, given the huge number of different experimental paradigms where MI is observable in various forms. Thus, (i) to (iii) should be seen as minimal set of necessary requirements. Next, we consider two first examples satisfying them.

Example 0.1 (Spike numbers) The following measure of MI in a single neuron is common in neurophysiology:

$$\mathrm{CRE}_{SP} = \frac{\mathrm{E}N_{VA} - \max\{\mathrm{E}N_V, \mathrm{E}N_A\}}{\max\{\mathrm{E}N_V, \mathrm{E}N_A\}} \times 100, \tag{0.1}$$

where $\mathrm{E}N_{VA}$ is the mean[3] (absolute) number of spikes in response to the crossmodal stimulus and $\mathrm{E}N_V, \mathrm{E}N_A$ denote the mean (absolute) numbers of spikes to the visual and auditory unisensory stimuli, respectively.[4] Thus,

[3] Note that we drop brackets in E[.] when there is no risk of confusion.
[4] Spike numbers are counted in a specified time interval and may or may not include spontaneous activity.

CRE_{SP} quantifies crossmodal enhancement/inhibition as the percentage difference between the response to a cross-modal pair VA and the largest response to one of its unisensory components, V or A.

Example 0.2 (Reaction time measure) An analogous measure for RTs is

$$\mathrm{CRE}_{RT} = \frac{\min\{\mathrm{E}RT_V, \mathrm{E}RT_A\} - \mathrm{E}RT_{VA}}{\min\{\mathrm{E}RT_V, \mathrm{E}RT_A\}} \times 100. \tag{0.2}$$

where $\mathrm{E}RT_{VA}$ is mean RT to an auditory-visual stimulus combination and $\min\{\mathrm{E}RT_V, \mathrm{E}RT_A\}$ is the faster of the unisensory mean RTs to the visual and auditory stimulus. Thus, CRE_{RT} expresses multisensory enhancement/inhibition as a proportion of the faster unisensory response. For example, $\mathrm{CRE}_{RT} = 10$ means that mean response time to the visual-auditory stimulus is 10% faster than the faster of the expected response times to unimodal visual and auditory stimuli.

0.3 Measures for the multisensory neuron response

0.3.1 Rules of multisensory integration

First systematic neuronal studies of MI, performed in the 1970s, focused on a midbrain structure, the cat *superior colliculus* (SC) (Meredith and Stein, 1983). Stein and colleagues showed that neurons in the deep layers of the SC are primary sites of multisensory convergence: if a visual-auditory stimulus combination is presented such that the visual stimulus is within its visual receptive field and the auditory stimulus is within its auditory receptive field, it will typically produce response enhancement, in the form of increased spike numbers, even when the stimuli are not found at the exact same spatial location. Likewise, response depression (inhibition) tends to occur if the visual stimulus is within its receptive field while the auditory stimulus is outside its receptive field. This has become known as the *spatial rule* of MI.

Similarly, changing the interval between auditory and visual stimulation can change enhancement to depression: presenting a visual stimulus 50 ms or 150 ms before the auditory ($V50A$ or $V150A$, for short) produced response enhancement, whereas longer intervals ($V300A$ or $A200V$) produced fewer impulses than a unisensory stimulus, i.e., depression (Meredith and Stein, 1983). The effect, termed *temporal rule* of MI, largely depends on the amount of overlap of the peak discharge periods of the neuron's unisensory responses. Later, these spatiotemporal rules of single neuron recordings have also been

observed in other species like the monkey, ferret, owl, guinea pig, rat, snake, and others.

A third major factor affecting MI is the efficacy of the component stimuli within the neuronal receptive fields. Response enhancement is found to be the greater the less effective the unisensory stimuli are. This rule of *inverse effectiveness* is most impressive when the unisensory stimulusintensities are below the threshold of eliciting any response from the neuron but in combination generate a reliable response.

More recently, a more nuanced function of unisensory signal strength and the temporal rule has been observed in cat SC (Miller et al., 2015). For each neuron, response magnitude (mean number of impulses per trial) to the visual (V) and the auditory stimuli (A) can be used to quantify the notion of *unisensory imbalance* (UI):

$$\mathrm{UI} = \frac{|\mathrm{E}N_A - \mathrm{E}N_V|}{\mathrm{E}N_A + \mathrm{E}N_V} \times 100. \tag{0.3}$$

UI quantifies the relative difference between the response magnitude to the visual and the auditory stimuli. It has a minimum of zero when the visual and auditory responses are of equal magnitude and a maximum of 100 when one of the responses is lacking.

In view of the above definition of crossmodal enhancement (Equation 0.1), increasing unisensory imbalance should not affect CRE_{SP}. However, across a wide range of response magnitude, increasing imbalance was found to be coupled with both a decrease in the multisensory response ($\mathrm{E}N_{VA}$) and in crossmodal enhancement CRE_{SP} (see Figure 0.1). Moreover, the order of arrival also mattered: when the unisensory response magnitudes were imbalanced, multisensory enhancement was maximized when stronger responses were advanced in time relative to weaker responses ("stronger first") and minimized when stronger responses were delayed ("stronger second") (for details, see Miller et al., 2015). Thus, only when the unisensory stimuli are "balanced", multisensory enhancement depends solely on their absolute temporal offset.

Still a different twist on the single-cell mechanism in SC has emerged from developmental findings. Since the early studies, it had been known that, just before and after birth, cat SC neurons are largely unresponsive to sensory stimulation and lack spontaneous activity. Successively, neurons start responding to tactile, then auditory, and finally visual stimulation. Besides unisensory neurons, multisensory neurons appear, but they do not yet show enhanced responses, instead they appear to act as a common conduit for different senses to reach the same motor output systems. These early studies

had shown that blocking an animal's multisensory experience, e.g. rearing cats with no visual stimulation at all, results in multisensory responses not stronger than the most effective component, suggesting CRE to be equal to zero. However, findings by Yu and colleagues (Yu et al., 2019) revealed that there exists competition between the senses in these "naïve" neurons: crossmodal stimuli, whether spatio-temporally disparate or not, can elicit inhibition in these neurons' responses. They conclude that the default mode of multisensory processing in SC is competition rather than absence of integration, and they develop a neurocomputational model consistent with this assumption. Thus, some form of MI (including competition) seems to occur at all stages of maturation, and the ability of enhanced (orienting) responses to crossmodal events increases over subsequent stages of development (Yu et al., 2019, p. 1374).

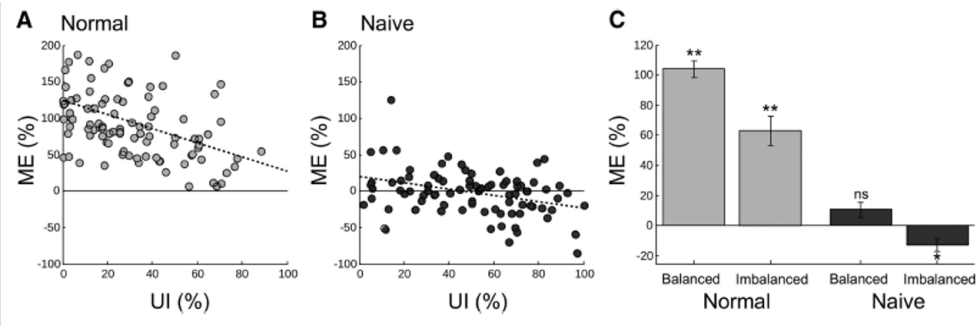

Figure 0.1 Relationships between multisensory responses (ME $\equiv$ CRE$_{SP}$) and unisensory imbalance (UI) in normal and naïve cohorts. Panel **A** Neurons from normally-reared animals produce their greatest response enhancements when the spatiotemporally concordant cues produced balanced unisensory responses: an inverse relationship between ME and UI (dotted line). Panel **B** Naïve SC neurons showed a similar inverse relationship between ME and UI, but even-balanced samples failed to produce significantly enhanced multisensory products, and imbalanced samples induced multisensory depression. Panel **C** Histograms summarizing the results. Vertical lines through the bars represent standard error (from Yu et al., 2019).

All the rules, sometimes referred to as *principles* of MI, discussed above have raised a discussion about whether, and in how far, they also determine multisensory behavior in humans and under more complex stimulus contexts. Before we follow up on these issues, we need to consider an aspect that proved particularly noteworthy in measuring MI.

0.3.2 Multisensory integration vs. probability summation

The fact that a multisensory neuron is responsive to multiple sensory modalities does not guarantee that it has actually engaged in integrating its multiple sensory inputs. Rather, it may simply respond to the most effective stimulus in a given trial, i.e. to the stimulus eliciting the strongest response.[5] In other words, it is possible that the response to a visual-auditory stimulus is simply determined by the larger of the responses to the modality-specific components, that is, by the the component that happens to elicit the higher absolute number of spikes in a given trial. Assuming random variation of the responses, such a mechanism is known as *probability summation* (PS).

In order to explore implications for how to measure MI in single neurons in the presence of PS, we first introduce some relevant statistical concepts. Only the case of facilitation will be discussed here, while the case of inhibition can be developed analogously. As before, the unisensory (visual, auditory) responses are conceived of as realizations of random variables N_V and N_A. We define distribution functions G_V and G_A, respectively:

$$\mathrm{P}\left[N_V \leq n_V\right] = G_V(n_V) \quad \text{and} \quad \mathrm{P}\left[N_A \leq n_A\right] = G_A(n_A),$$

with n_V and n_A taking integer values $0, 1, \ldots$. For the bisensory condition, we assume a distribution function G_{VA} exists such that

$$\mathrm{P}\left[N_{VA} \leq n\right] = G_{VA}(n),$$

with $n = 0, 1, \ldots$. Thus, N_V, N_A, and N_{VA} are random variables whose realizations (samples) are observed in the experiment under to-be-specified conditions.

0.3.2.1 Probability summation (PS) in spike numbers

For clarity, the three assumptions underlying the concept of PS in this multisensory context will be stated in detail. The first assumption refers to the observation that realizations of the random variables N_V and N_A are collected under different stimulus conditions (visual vs. auditory) and, thus, occur in distinct probability spaces. A–priori, there is no prescribed way how to combine them. In particular, any assumption about stochastic (in-)dependence between N_V and N_A is meaningless. However, one can postulate a *stochastic coupling*[6] of the two random variables.

[5] As Stein and colleagues (Stein et al., 2009, p. 114) have put it, "*At the time of the early physiology studies in the 1980s, it was considered possible that these neurons only represented a common route by which independent inputs from a variety of senses could gain access to the same motor apparatus in generating behavior (e.g., possibly employing a "winner-take-all" algorithm).*"

[6] See Colonius (2016) for an introduction to that concept in this context.

Assumption 1: There exists a random vector $(\tilde{N}_V, \tilde{N}_A)$ *with a joint distribution* $\tilde{H}_{VA}$*,*

$$\tilde{H}_{VA}(n_V, n_A) = \mathrm{P}\,[\tilde{N}_V \leq n_V, \tilde{N}_A \leq n_A].$$

Assuming the existence of $\tilde{H}_{VA}$ amounts to a *coupling* of the random variables $\tilde{N}_V$ and $\tilde{N}_A$, which is always possible. Of course, we want $\tilde{N}_V$ and $\tilde{N}_A$ to be a "copy" of N_V and N_A in the following sense:

Assumption 2: The marginal distributions of $\tilde{H}_{VA}(n_V, n_A)$ *are equal to* G_V *and* G_A*, respectively:*

$$\tilde{H}_{VA}(n_V, \infty) = G_V(n_V) \quad \textit{and} \quad \tilde{H}_{VA}(\infty, n_A) = G_A(n_A).$$

This important restriction, equating the marginals to the observable unisensory response distributions, is often called "context invariance".

Note that we have not assumed a specific form for $\tilde{H}_{VA}$. In fact, we are only interested in the values on the diagonal, $\tilde{H}_{VA}(n, n)$. For $n = 0, 1, \ldots$, we write

$$\begin{aligned}\tilde{H}_{VA}(n, n) &= \mathrm{P}\,[\{\tilde{N}_V \leq n\} \cap \{\tilde{N}_A \leq n\}] \\ &= \mathrm{P}\,[\max\{\tilde{N}_V, \tilde{N}_A\} \leq n] \\ &\equiv \tilde{G}_{VA}(n).\end{aligned}$$

The third assumption specifies the probability mechanism proper:

Assumption 3: For $n = 0, 1, \ldots$

$$G_{VA}(n) = \tilde{G}_{VA}(n), \tag{0.4}$$

that is, the observable crossmodal responses are the result of taking the maximum of the unisensory responses.

It is always possible to construct *some* bivariate distribution $\tilde{H}_{VA}(n_V, n_A)$, e.g., by assuming stochastic independence:

$$\tilde{H}_{VA}(n_V, n_A) = \mathrm{P}\,[\tilde{N}_V \leq n_V]\,\mathrm{P}\,[\tilde{N}_A \leq n_A],$$

which implies the empirically testable hypothesis

$$G_{VA}(n) = \tilde{G}_{VA}(n) = G_V(n)\,G_A(n)$$

for $n = 0, 1, \ldots$, under context invariance (*Assumption 2*).

In general, however, it is not obvious how *Assumption 3* should be tested. Stochastic independence, while convenient, may not be the most judicious choice, as will be argued below.

0.3.3 Measures of MI under PS hypothesis

It is straightforward to compare observed responses with those predicted by PS: one has to gauge the difference between the means (expected values) associated with G_{VA} and $\tilde{G}_{VA}$, that is $\mathrm{E}N_{VA}$ and $\mathrm{E}\max\{N_V, N_A\}$, respectively. The common measure of MI based on spike counts introduced in Example 0.1,

$$\mathrm{CRE}_{SP} = \frac{\mathrm{E}N_{VA} - \max\{\mathrm{E}N_V, \mathrm{E}N_A\}}{\max\{\mathrm{E}N_V, \mathrm{E}N_A\}} \times 100, \tag{0.5}$$

is then replaced by

$$\mathrm{CRE}^*_{SP} = \frac{\mathrm{E}N_{VA} - \mathrm{E}\max\{N_V, N_A\}}{\mathrm{E}\max\{N_V, N_A\}} \times 100. \tag{0.6}$$

Note that *Assumption 2* permits us to write measure CRE^*_{SP} with N_V, N_A instead of $\tilde{N}_V, \tilde{N}_A$. By a well-known statistics result (*Jensen's Inequality*, e.g. Ross, 1996),

$$\max\{\mathrm{E}N_V, \mathrm{E}N_A\} \leq \mathrm{E}\max\{N_V, N_A\}$$

always holds, obviously implying

$$\mathrm{CRE}^*_{SP} \leq \mathrm{CRE}_{SP}. \tag{0.7}$$

This inequality reveals an important consequence: in order to assess "true" MI, that is, over and above the effect of PS, the criterion *mean* number of spikes observed ($\mathrm{E}N_{AV}$) has to be larger than the mean taking PS into account.

0.3.3.1 Effects of unisensory imbalance

The move from CRE_{SP} to CRE^*_{SP} opens up the possibility to probe effects of unisensory imbalance mentioned above (Equation 0.3),

$$\mathrm{UI} = \frac{|\mathrm{E}N_A - \mathrm{E}N_V|}{\mathrm{E}N_A + \mathrm{E}N_V}.$$

Note that only the maximum of $\mathrm{E}N_A$ and $\mathrm{E}N_V$ enters into CRE_{SP}, so that varying imbalance has no effect on that index. In contrast, computing $\mathrm{E}\max\{N_V, N_A\}$ involves the distribution of both variables, N_A and N_V, and it is easy to find instances where CRE^*_{SP} depends on both $\mathrm{E}N_A$ and $\mathrm{E}N_V$ simultaneously (see, e.g. Colonius and Diederich (2017) for an example with Poisson-distributed spike counts).

0.3.3.2 Towards an optimal measure of MI

Inequality (0.7) holds without assuming a specific distribution for $\tilde{G}_{VA}$. While stochastic independence between N_V and N_A is typically taken for granted in computing the value of $\mathrm{E}\max\{N_V, N_A\}$, it turns out that it is not the most conservative choice possible.[7] To demonstrate, we recall (without proof) a classic result from statistics (Fréchet, 1951) about upper and lower bounds for arbitrary distributions, here applied to $\tilde{H}_{VA}$.

Lemma 0.3 (Fréchet inequalities) *For $m, n = 0, 1, \ldots$, let $\tilde{H}_{VA}(m, n) = P(\tilde{N}_V \leq m, \tilde{N}_A \leq n)$ be a bivariate distribution with marginals $\tilde{G}_V(m), \tilde{G}_A(n)$, respectively. Then,*

$$\max\{0, \tilde{G}_V(m) + \tilde{G}_A(n) - 1\} \leq \tilde{H}_{VA}(m, n) \leq \min\{\tilde{G}_V(m), \tilde{G}_A(n)\}.$$

The upper and lower bound in the lemma represent bivariate distributions as well, with the same marginals as $\tilde{H}_{VA}(m, n)$ but possessing maximal positive, respectively negative, dependence between $\tilde{N}_V$ and $\tilde{N}_A$ (e.g., Joe, 1997). Setting $m = n$, we denote the lower bound with maximal negative dependence by $\tilde{G}^{(-)}_{VA}(n)$. Then,

$$\tilde{G}^{(-)}_{VA}(n) \equiv \max\{0, \tilde{G}_V(n) + \tilde{G}_A(n) - 1\} \leq \tilde{G}_{VA}(n) \tag{0.8}$$

for $n = 0, 1, \ldots$.

Importantly, maximal negative dependence between $\tilde{N}_V$ and $\tilde{N}_A$ maximizes the expected value of $\mathrm{E}\max\{N_V, N_A\}$:

Lemma 0.4 *Let $E^{(-)}\max\{\tilde{N}_V, \tilde{N}_A\}$ be the expected value of $\max\{\tilde{N}_V, \tilde{N}_A\}$ under bivariate distribution $\max\{0, \tilde{G}_V(m) + \tilde{G}_A(n) - 1\}$; then*

$$E\max\{\tilde{N}_V, \tilde{N}_A\} \leq E^{(-)}\max\{\tilde{N}_V, \tilde{N}_A\}$$

under any bivariate distribution $\tilde{H}_{VA}(m, n)$ for $E\max\{\tilde{N}_V, \tilde{N}_A\}$.

This can be shown as follows. Rewriting Equation (0.8) as

$$1 - \tilde{G}_{VA}(n) \leq 1 - \tilde{G}^{(-)}_{VA}(n)$$

and summing over all n yields

$$\mathrm{E}\max\{N_V, N_A\} \equiv \sum_{n=0}^{\infty}[1 - \tilde{G}_{VA}(n) \leq \sum_{n=0}^{\infty}[1 - \tilde{G}^{(-)}_{VA}(n) \equiv \mathrm{E}^{(-)}\max\{N_V, N_A\}.$$

The upshot of Lemma 0.4 is that an optimal choice for defining CRE^{*}_{SP} (Equation 0.6) is to insert $\mathrm{E}^{(-)}\max\{N_V, N_A\}$:

[7] Here, 'conservative' means that one wants to avoid claiming MI to hold when, in reality, it does not.

Definition 0.5 The measure of MI taking into account PS with maximal negative dependence between the unisensory responses is

$$\mathrm{CRE}_{SP}^{max} = \frac{\mathrm{E}N_{VA} - \mathrm{E}^{(-)} \max\{N_V, N_A\}}{\mathrm{E}^{(-)} \max\{N_V, N_A\}} \times 100. \tag{0.9}$$

Note that it is not claimed here that a multisensory neuron actually operates under this extreme negative dependency rule. As long as PS is considered as possible alternative to "true" MI, however, some specification of the stochastic relation between the unisensory responses has to be made in CRE_{SP}^{*}. Assuming maximal negative dependency is simply the most efficient way to hedge against a "false alarm", that is, declaring true MI while enhancement may simply be a product of PS. Whenever there is empirical or theoretical evidence in favor of some other form of dependence, e.g. stochastic independence, this could be used to modify the benchmark appropriately.

Because, in general, the new measure is more restrictive than the traditional CRE measure, many neurons previously categorized as "multisensory" may lose that property. The purpose of the new measure corresponds to that of the traditional measure: given a fixed statistical criterion, one may categorize a single neuron as either being "multisensory" or not. It is of course possible that a neuron actually "truly" integrates the unimodal activations but still does not meet the criterion set by maximal negative PS. However, as long as one has no direct insight into the integration mechanism, an alternative interpretation in terms of PS simply cannot be ruled out.

0.3.3.3 Example application of CRE_{SP}^{max}

Estimating $\mathrm{E}\max\{\tilde{N}_V, \tilde{N}_A\}$ from sample data is straightforward. Without going into detail, the procedure is as follows. We have two samples of numbers of spikes from each modality of size n_v and n_a, say, and assume $n_v = n_a$. Under the stochastic independence version of SP, the number of spikes occurring in trial $i, i = 1, \ldots, n_v$ is randomly paired with the number of spikes in trial $j, j = 1, \ldots, n_a$ (without replacement). The maximum in each pair is determined and the average of the maxima yields an estimate of $\mathrm{E}\max\{\tilde{N}_V, \tilde{N}_A\}$.

Under maximal negative dependence of PS, trial i with the largest number of spikes is paired with trial j with the smallest number of spikes, the second largest i is paired with the second lowest j, and so on (method of 'antithetic variables'), and the average of the maxima is again computed as estimate of $\mathrm{E}\max\{\tilde{N}_V, \tilde{N}_A\}$. If the unisensory samples are of different sizes, some replacement procedure could be applied. In an illustrative sample of cat SC

neurons[8], Colonius and Diederich (2017) showed that there was a significant decrease from CRE_{SP} to CRE_{SP}^{max} in 24 out of 27 recording blocks collected from 20 neurons. Whether or not the label "multisensory" is actually lost for some neurons, however, depends on criteria of the statistical test comparing the sample means (see Figure 0.2).

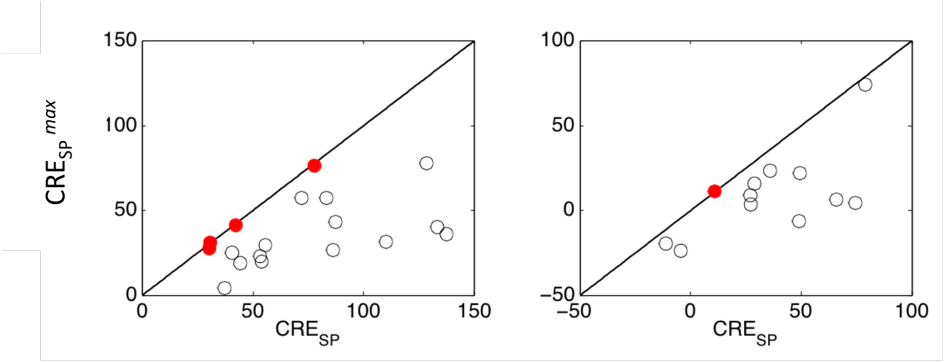

Figure 0.2 Pairs of sample estimates of $(\mathrm{CRE}_{SP}, \mathrm{CRE}_{SP}^{max})$ based on 27 recording blocks (15 stimulus presentations in each block). In the left-hand panel spontaneous activity was included, in the right-hand panel it has been removed.. Filled circles indicate no significant difference between CRE_{SP} and CRE_{SP}^{max}, based on bootstrap confidence intervals (N = 10,000, $\alpha = 0.05$). Thus, each open circle refers to a recording where the label multisensory may be lost when applying measure CRE_{SP}^{max}. There were 4 out of 27 cases with no significant difference between both measures (left panel), after spontaneous activity was removed, only 1 out of 19 cases was not significant (right panel) (from Colonius and Diederich, 2017).

0.4 Measures based on response speed

The earliest observations of MI effects have likely been reported in the context of measuring the speed and accuracy of responses to crossmodal stimuli at the beginning of the 20th century (Welch and Warren, 1986, for a review). In a typical paradigm, participants are instructed to respond via button press as soon as a signal of any modality occurs (*redundant signals paradigm*[9]). It is to be distinguished from a related paradigm, often called *focused attention paradigm*; in the latter, one modality is designated as 'target' modality, the other as 'distractor' modality, and participants are instructed

[8] Data provided by the lab of Mark Wallace (personal communication).
[9] Also known as *divided attention paradigm*.

to respond only to signals from the target modality (mostly, visual) but not to distractor signals. The two paradigms demand separate treatments for the measurement of MI.

Note that erroneous responses should also be defined differently for the two paradigms, but we will first ignore errors entirely since they are often kept at a negligible rate in the experiments. Accuracy measures are discussed later.

0.4.1 MI measures in redundant signals paradigms

In general, bisensory, in particular visual-auditory, stimulation results in smaller mean RT compared to unisensory stimulation, and responses to trisensory stimulation (often visual, auditory, and tactile) are faster on average than to bisensory stimulation. The magnitude of the speed-up depends on the specifics of the experiment, in particular the intensity of the different modalities and their temporal configuration. For visual-auditory presentations, the greatest effect is typically found when the visual stimulus precedes the auditory by an interval that equals the difference between the unisensory mean RTs.

Hence the MI measure for RTs introduced in Equation (0.2) should be augmented to include stimulus onset asynchrony (SOA), denoted as τ:

$$\mathrm{CRE}_{RT,\tau} = \frac{\min\{\mathrm{E}RT_V, \mathrm{E}RT_A + \tau\} - \mathrm{E}RT_{V\tau A}}{\min\{\mathrm{E}RT_V, \mathrm{E}RT_A + \tau\}} \times 100, \tag{0.10}$$

where $RT_{V\tau A}$ is the RT to a visual-auditory stimulus combination with the visual preceding the auditory by τ [ms]; thus, the maximum of $\mathrm{CRE}_{RT,\tau}$ would be expected[10] for $\tau = \mathrm{E}RT_V - \mathrm{E}RT_A$.

For trisensory stimulus contexts (VAT), the analogous measure is

$$\mathrm{CRE}_{RT,\tau_1\tau_2} = \frac{\min\{\mathrm{E}RT_V, \mathrm{E}RT_A + \tau_1, \mathrm{E}RT_T + \tau_1 + \tau_2\} - \mathrm{E}RT_{V\tau_1 A\tau_2 T}}{\min\{\mathrm{E}RT_V, \mathrm{E}RT_A + \tau_1, \mathrm{E}RT_T + \tau_1 + \tau_2\}} \times 100, \tag{0.11}$$

where $RT_{V\tau_1 A\tau_2 T}$ is the RT to a visual-auditory-tactile stimulus combination with the visual preceding the auditory by τ_1 [ms] and the auditory preceding the tactile by τ_2 [ms].

Note that adding a third modality increases the possible measures of response enhancement: trisensory response speed may now also be compared with the speed of any bisensory combination, e.g. $V\tau_1 A\tau_2 T$ with $V\tau_1 A$ or $A\tau_2 T$, as long as these different combinations have been presented in the

[10] Visual RTs tend to be slower than auditory RTs at comparable intensity levels.

experiment. For example,

$$\mathrm{CRE}_{RT,(\tau_1)\tau_2} = \frac{\mathrm{E}RT_{V\tau_1 A} - \mathrm{E}RT_{V\tau_1 A\tau_2 T}}{\mathrm{E}RT_{V\tau_1 A}} \times 100, \tag{0.12}$$

measuring the additional multisensory effect of a tactile stimulus, presented τ_2 [ms] later, on the speed of a visual-auditory combination.

0.4.2 Probability summation in the redundant signals paradigm

None of the RT measures of MI considered so far takes the PS hypothesis into account. In this context, the hypothesis amounts to postulating the so-called *race model* and as such, arguably, represents the most widely known version of PS in multisensory research. The idea is that, e.g., a visual-auditory stimulus combination triggers random visual and auditory processing times such that the observed RT equals the minimum of the two, i.e. the 'winner of the race'.

Usually, RTs are assumed to comprise some additive components, like motor preparation and execution. To simplify the discussion, we neglect this distinction here. Observed samples from random variables, denoted as T_V, T_A, and T_{VA} represent RTs obtained in unisensory visual, auditory, and bisensory trials, respectively. Thus, we equate realizations of T_V, T_A, and T_{VA} with the observable RT under these conditions.

We define underlying distribution functions F_V and F_A, respectively:

$$\mathrm{P}\left[T_V \leq t_V\right] = F_V(t_V) \quad \text{and} \quad \mathrm{P}\left[T_A \leq t_A\right] = F_A(t_A),$$

with T_V and T_A taking on nonnegative real numbers. For the bisensory context, we assume a distribution function F_{VA} such that

$$\mathrm{P}\left[T_{VA} \leq t\right] = F_{VA}(t),$$

with $t \geq 0$. Hence, T_V, T_A, and T_{VA} are random variables whose realizations are observed in an experiment under to-be-specified conditions.

0.4.2.1 Probability summation (PS) in reaction times

The exact definition of PS follows in close analogy to the one given for spike numbers in the previous section:

Assumption 1: There exists a random vector $(\tilde{T}_V, \tilde{T}_A)$ with a joint distribution $\tilde{K}_{VA}$,

$$\tilde{K}_{VA}(t_V, t_A) = \mathrm{P}\left[\tilde{T}_V \leq t_V, \tilde{T}_A \leq t_A\right].$$

Assuming the existence of $\tilde{K}_{VA}$ amounts again to a *coupling* of the random variables $\tilde{T}_V$ and $\tilde{T}_A$, which is always possible. Of course, we want $\tilde{T}_V$ and $\tilde{T}_A$ to be a "copy" of T_V and T_A in the following sense:

Assumption 2: The marginal distributions of $\tilde{K}_{VA}(t_V, t_A)$ are equal to F_V and F_A, respectively:

$$\tilde{K}_{VA}(t_V, \infty) = F_V(t_V) \quad \textit{and} \quad \tilde{K}_{VA}(\infty, t_A) = F_A(t_A).$$

Thus, "context invariance" is postulated for RT distributions as well. It follows that for $t \geq 0$,

$$\begin{aligned} \tilde{K}_{VA}(t,t) &= \mathrm{P}\left[\{\tilde{T}_V \leq t\} \cap \{\tilde{T}_A \leq t\}\right] \\ &= \mathrm{P}\left[\max\{\tilde{T}_V, \tilde{T}_A\} \leq t\right] \\ &\equiv \tilde{F}_{VA}(t). \end{aligned}$$

Assumption 3: For $t \geq 0$

$$F_{VA}(t) = \tilde{F}_V(t) + \tilde{F}_A(t) - \tilde{F}_{VA}(t). \tag{0.13}$$

This second assumption is the central one again, implying that the observable crossmodal RTs result from taking the minimum of the unisensory RTs (race model).

It is always possible to construct some bivariate distribution $\tilde{K}_{VA}(t_V, t_A)$, e.g., by assuming stochastic independence:

$$\begin{aligned} \tilde{K}_{VA}(t_V, t_A) &= \mathrm{P}\left[\tilde{T}_V \leq t_V\right] \mathrm{P}\left[\tilde{T}_A \leq t_A\right] \\ &= F_V(t_V)\, F_A(t_A) \qquad \text{by } \textit{Assumption 2}, \end{aligned}$$

implying the special case of "independent race model"

$$F_{VA}(t) = 1 - (1 - F_V(t))(1 - F_A(t)). \tag{0.14}$$

The PS hypothesis has been studied as a possible non-parametric model for RTs in the redundant signals paradigm. Being equivalent to the 'race model', it predicts a specific relation between the distribution functions for bisensory and the unisensory conditions:

$$\begin{aligned} F_{VA}(t) &= \tilde{F}_V(t) + \tilde{F}_A(t) - \tilde{F}_{VA}(t) \qquad \text{by } \textit{Assumption 3} \\ &= F_V(t) + F_A(t) - \tilde{F}_{VA}(t) \qquad \text{by } \textit{Assumption 2} \\ &\leq F_V(t) + F_A(t). \end{aligned} \tag{0.15}$$

Inequality (0.15) is a simple version of *Boole's inequality* and has been called "race-model inequality" (RMI) in this context. Testing it has become routine

in a vast number of empirical studies, using a variety of different statistical procedures[11]. Note that the right-hand side of RMI approaches 2 for t going to infinity, so it can be replaced by $\min\{F_V(t) + F_A(t), 1\}$. Typically, RMI tends to be violated for t not too large.

0.4.3 Measures of MI in redundant signals paradigms under PS

In addition to testing the race model, a quantitative measure of the degree of RMI violation has been proposed. The latter turns out to be the basis of a measure of MI in redundant signal experiments.

We define a function $R_{VA}(t)$, for $t \geq 0$,

$$R_{VA}(t) \equiv F_{VA}(t) - \min\{F_V(t) + F_A(t), 1\}. \tag{0.16}$$

Hence, values of t with $R_{VA}(t) > 0$ indicate a violation of RMI, whereas values of t with $R_{VA}(t) \leq 0$ are compatible with the race model. The positive part of the area between $F_{VA}(t)$ and $\min\{F_V(t) + F_A(t), 1\}$ is often taken as measure of the amount of RMI violation. Integrating $R_{VA}(t)$ results in a convenient interpretation as MI measure. First, observe that

$$\begin{aligned} R_{VA}(t) &= F_{VA}(t) - \min\{F_V(t) + F_A(t), 1\} \\ &= 1 - \min\{F_V(t) + F_A(t), 1\} - [1 - F_{VA}(t)] \\ &= \max\{1 - F_V(t) - F_A(t), 0\} - [1 - F_{VA}(t). \end{aligned}$$

Integrating yields

$$\begin{aligned} \int_0^\infty R_{VA}(t)\, dt &= \int_0^\infty \max\{1 - F_V(t) - F_A(t), 0\}\, dt - \int_0^\infty [1 - F_{VA}(t)\, dt \\ &= \mathrm{E}^{(-)} \min\{T_V, T_A\} - \mathrm{E}\{T_{VA}\}, \end{aligned}$$

where $\mathrm{E}^{(-)} \min\{T_V, T_A\}$ denotes mean RT predicted by a race model with maximal negative dependence between the latencies T_V and T_A. This leads to a modified version of $\mathrm{CRE}_{RT,\tau}$ (see Equation 0.10) accounting for PS,

$$\mathrm{CRE}^{min}_{RT,\tau} = \frac{\mathrm{E}^{(-)} \min\{RT_V, RT_A + \tau\} - \mathrm{E}RT_{V\tau A}}{\mathrm{E}^{(-)} \min\{RT_V, RT_A + \tau\}} \times 100, \tag{0.17}$$

where T_V, T_A are identified with RT_V, RT_A, respectively.

[11] Sometimes, the 'independent' version of the inequality is tested, $F_{VA}(t) \leq F_V(t) + F_A(t) - F_V(t)F_A(t)$, but violations of this inequality would only rule out the special case of a stochastically independent race.

0.4.4 MI measures in focused attention paradigms

Let us assume a stimulus from the visual modality is the target. The task is to respond to the occurrence of the target, via button press, while ignoring an auditory stimulus ('distractor') presented in spatio-temporal proximity. In a frequent variant, the required response is to execute an eye movement towards a target that occurs at a randomized spatial position in the visual field, with saccadic RT and/or accuracy of the trajectory/landing position being recorded. In all cases, MI is measured by how much the response to the target is modulated by the presence of a distractor. For RTs, a simple adaption of the CRE measure in the redundant paradigm results in

$$\mathrm{CRE}_{RT} = \frac{\mathrm{E}RT_V - \mathrm{E}RT_{VA}}{\mathrm{E}RT_V} \times 100. \tag{0.18}$$

The amount and direction (facilitation vs. inhibition) of CRE_{RT} depends on a host of experimental conditions. Because visual and auditory stimuli activate visuomotor neurons in superior colliculus (SC) thereby eliciting goal-directed eye movements, many studies of MI have focused on gaze behavior, in particular saccadic reaction time.[12]

While the temporal and spatial rules of MI are, in general, consistent with findings in the redundant signals task, effects of the role of localizability of the auditory distractor has found special attention in eye movement experiments. Specifically, when target and distractor are presented at the same position (e.g., both above or below fixation point), SRTs are faster than when they are presented at opposite positions (e.g., target above, distractor below fixation point). However, this effect disappears when localization of the auditory stimulus is made more difficult, e.g. by increasing the level of a background noise. Hence, the *perceived* rather than the physical distance between target and distractor controls the MI effect (Colonius et al., 2009).

0.5 MI measures based on accuracy

Next, we discuss MI measures based on accuracy. These measures turn up in a variety of multisensory tasks, including detection, discrimination, recognition, and identification. We will not be able to cover all of them, but rather focus on a few important aspects.

[12] We limit the presentation here to SRTs, MI measures involving other aspects of eye movements are similarly obtainable.

0.5.1 MI measures based on detection accuracy

Let p_V, p_A, and p_{VA} denote the probability of responding "Yes" to the question of whether a visual, auditory, or combined visual-auditory stimulus has been presented, respectively. In analogy to CRE measures of response speed in the redundant signals task, we define *crossmodal detection rate* as

$$\mathrm{CRE}_{DR} = \frac{p_{VA} - \max\{p_V, p_A\}}{\max\{p_V, p_A\}} \times 100. \tag{0.19}$$

Typically, the probability of a "Yes" response will primarily depend on stimulus intensity. If at least one of the unisensory stimuli is clearly detectable (i.e., p_A or p_V close to one), p_{VA} will also be close to one, and so crossmodal detection rate will be close to zero. If intensity is low or, equivalently, the level of noise during presentation is (moderately) high, determining the likelihood to respond "Yes" is not straightforward: the participant may have a tendency to guess and/or may have an internal criterion for responding "Yes" or "No" which leads us to the realm of signal detection theory (SDT) (Green and Swets, 1974).

In the terminology of SDT, it is not sufficient to compare the crossmodal *hit rate* (probability to say "Yes" when the stimulus is presented) with the unisensory hit rates because increasing the hit rate often goes along with increasing the *false-alarm rate* (probability to say "Yes" when no stimulus is presented) as well. Assuming the standard equal-variance Gaussian distribution model of SDT, CRE_{DR} can be replaced by inserting the corresponding d-prime measures,

$$\mathrm{CRE}_{SDT} = \frac{d'_{VA} - \max\{d'_V, d'_A\}}{\max\{d'_V, d'_A\}} \times 100. \tag{0.20}$$

This measure assesses the relative amount of sensitivity increase in the visual-auditory condition compared to the best unisensory condition, while separating sensitivity from possible biases to respond "Yes" or "No" in each condition. An analogous definition for the focused attention task is obvious.

Measure CRE_{SDT} tests against a benchmark where the observer simply ignores the less detectable modality. However, it is also possible to modify CRE_{SDT} such that a PS strategy is taken into account. Let us assume that an observer sets two criteria, λ_V and λ_A, and a "Yes" response is given if at least one of the criteria is exceeded. Under stochastic independence, the probabilities of *misses* (1 minus probability of a hit) and *correct rejections* (1 minus probability of a false alarm) are the product of their modality components. Writing f_V, f_A, h_V, h_A, and f_{VA}, h_{VA} for the false-alarm and

hit rates for the unisensory and bisensory conditions, respectively, we get

$$f_{VA} = 1 - (1 - f_V)(1 - f_A) = 1 - \Phi(\lambda_V)\Phi(\lambda_A)$$
$$h_{VA} = 1 - (1 - h_V)(1 - h_A) = 1 - \Phi(\lambda_V - d'_V)\Phi(\lambda_A - d'_A),$$

with Φ denoting the standard Gaussian distribution function. From this we can compute the visual-auditory sensitivity under the PS strategy,

$$d'^{PS}_{VA} = \Phi^{-1}(h_{VA}) - \Phi^{-1}(f_{VA}).$$

Inserting into expression (0.20) results in a modified measure of response enhancement gauging against PS,

$$\mathrm{CRE}^{PS}_{SDT} = \frac{d'_{VA} - d'^{PS}_{VA}}{d'^{PS}_{VA}} \times 100. \tag{0.21}$$

Besides the PS notion, numerous alternative models on how unisensory detection accuracy is combined into a bisensory one have been discussed in the literature (see Jones, 2016, for a recent tutorial). Finally, when there is empirical evidence against the equal-variance assumption of SDT, alternative measures, like the area under the operating characteristic, may be considered instead of *d*-prime values (see, e.g., Lovelace et al., 2003, for a focused-attention example).

0.5.2 Measures for audiovisual speech identification

Arguably, one of the most thoroughly studied line of multisensory research is the identification of speech in an audiovisual paradigm. In typical audiovisual speech identification (or recognition) tests, listeners are presented with audio materials like syllables, words, phrases, or sentences along with a video of a speaker's face acquired at the same time as the audio materials. Commonly, speech heard in noise (often, talker babble noise at different levels) can be more accurately identified or recognized when the participant sees a speaker's articulating face or lip movements.

However, there still seems to be considerable controversy with respect to the source of this audiovisual advantage. According to several studies, when hearing-impaired individuals, or different age groups, are compared with respect to the amount of audiovisual benefit, one finds large differences across individuals or groups. Notably, these differences are often found to persist even when differing unisensory auditory or visual speech recognition performance levels are taken into account. Thus, besides lipreading ability and auditory encoding ability, an ability to integrate auditory and visual information should be assessed in order to explain audiovisual performance

(Grant, 2002). In contrast, it is also held that an audiovisual speech signal represents a more robust representation of any given word because, first, simultaneous auditory and visual speech signals provide complementary information: vision contributes clues about some aspects of the speech event that are hard to hear and which may depend on the shape and contour of the lower face being clearly visible. Second, reinforcing information may be provided by the temporal congruence between amplitude fluctuations in the auditory signal and mouth opening and closing in the visual signal. That is, when the auditory signal gets louder, the visible mouth and jaw tend to be opening; when the signal gets softer, the mouth and jaw tend to be closing (see Tye-Murray et al., 2016).

0.5.2.1 Measures of response enhancement and superadditivity

Without subscribing to a specific source of the audiovisual advantage, ad-hoc measures of enhancement have been developed. Letting p_{AV} denote the probability[13] of correctly identifying words in the audiovisual condition and p_V, p_A the corresponding probability in the vision-only and auditory-only condition, respectively, one defines *visual enhancement* (VE) as

$$\mathrm{VE} = \frac{p_{AV} - p_A}{1 - p_A}. \tag{0.22}$$

Thus, VE represents the amount of benefit afforded by the addition of the visual channel of speech, normalized for the amount of possible improvement. Analogously, one defines *auditory enhancement* (AE) as

$$\mathrm{AE} = \frac{p_{AV} - p_V}{1 - p_V}. \tag{0.23}$$

Thus, AE represents the amount of benefit afforded by the addition of the auditory channel of speech, again normalized for the amount of possible improvement.

Although these enhancement measures do not seem controversial, some criticism has been raised against them. First, whereas there is broad empirical support for the principle of inverse effectiveness (Section 0.3.1) being valid in audiovisual speech performance, the normalization involved in calculating AE biases against finding results consistent with it. Specifically, among listeners with equivalent improvement (i.e. equal numerators), AE will be lower for those who made more lipreading errors, inconsistent with the principle (as pointed out by Tye-Murray et al., 2010, p. 639).

Second, a more sweeping argument was recently made by Dias et al. (2021)

[13] Note that p_V, p_A, and p_{AV} here are not the same as in the previous section on detection, but no confusion should arise.

studying the mean proportion of correctly identified words for two different age groups. Consistent with previous research, they found p_V and p_A to decline with age and to correlate positively with each other, but p_{AV} did not significantly differ between age groups. Importantly, they did not find VE and AE to exhibit any age effects. Dias and colleagues offer the following explanation, after defining "superadditivity" p_{sAV} as

$$p_{sAV} = p_{AV} - (p_A + p_V). \tag{0.24}$$

Rewriting the expressions for VE and AE yields

$$\text{VE} = \frac{p_{AV} - p_A}{1 - p_A} = \frac{p_V + p_{sAV}}{1 - p_A}$$

and

$$\text{AE} = \frac{p_{AV} - p_V}{1 - p_V} = \frac{p_A + p_{sAV}}{1 - p_V}.$$

The superadditivity term occurring in both VE and AE explains the positive correlation; moreover, the authors argue, the absence of an age effect is due to the declining values of p_V and p_A with age canceling an alleged increase of superadditivity, p_{sAV}, also with age.[14]

0.5.2.2 Measures derived from modeling audiovisual speech identification

Different models of auditory-visual speech integration have been proposed. They often predict "optimal" performance in the bisensory condition given the information extracted in the unimodal conditions separately (e.g., for nonsense syllables, words, or sentences), thereby providing quantitative measures of *integration efficiency* (IE).

The simplest one is a model representing a PS version of crossmodal detection rate CRE_{DR} (Equation (0.19). Assuming independent PS for auditory and visual performance, the probability p^I_{AV} to recognize an item in the audiovisual condition equals

$$p^I_{AV} = 1 - (1 - p_A) \times (1 - p_V) = p_A + p_V - p_A \times p_V.$$

From this, integration efficiency is defined as (e.g. Tye-Murray et al., 2007)

$$\text{IE}^I = \frac{p^{obs}_{AV} - p^I_{AV}}{1 - p^I_{AV}}, \tag{0.25}$$

where p^{obs}_{AV} is the observed probability in the audiovisual condition. Integration efficiency measured this way has often been found to be positive, but

[14] Dias et al. (2021) use notation AO, VO, and AV instead of probabilities; see the paper for details of their exhaustive statistical analyses.

some recent findings support the PS model as well (van de Rijt et al., 2019) implying zero integration efficiency.

A prominent model for audiovisual speech identification is Massaro's *fuzzy logical model of perception* (FLMP) with an optimal integration rule equivalent to Bayes' theorem (see Massaro and Cohen, 2000).

Prelabeling model of integration (PRE). Another widely known model is Braida's PRE model (Braida, 1991) where each response R_j corresponds to a point in a D-dimensional Euclidean vector space of stimulus attributes (cue vectors) referred to as *prototypes*. Each presentation of a stimulus i generates a D-dimensional vector of cues X in the same space following a multivariate normal distribution with independent components, unit variance, and a given mean S_i not necessarily identical to the prototype corresponding to R_i. According to a decision rule of multidimensional signal detection theory, the subject responds R_j if and only if the (Euclidean) distance of X to the prototype of R_j is smaller than the distance to any other prototype. The prototype locations are assumed to reflect response bias effects, whereas the subject's sensitivity in discriminating stimulus i from stimulus j, d-prime value $d'(i,j)$, is given by the Euclidean distance between S_i and S_j. The model parameters, i.e., the components of vectors S_i and R_i, are estimated iteratively through non-metric multidimensional scaling by comparing observed and predicted confusion matrices. The decision space for the AV condition is assumed to be the Cartesian product of the space for the A condition and the space for the V condition. A subject's sensitivity in the AV condition can be shown to be related to the unimodal sensitivities by

$$d'_{AV}(i,j) = \sqrt{d'_A(i,j)^2 + d'_V(i,j)^2}. \tag{0.26}$$

An IE measure is then defined by taking the ratio between the obtained and predicted d'_{AV} scores:

$$\mathrm{IE}^{PRE} = \frac{d'_{AV}(obs)}{d'_{AV}(pred)}. \tag{0.27}$$

Note that perfect integration need not be associated with high overall AV performance: If a participant has very bad hearing or is a very poor speech reader, it is unlikely that they will achieve a high AV score. Nevertheless, a subject may still integrate the available A and V cues in a nearly optimal manner, and if so, the integration efficiency measure should be near unity (see Figure 0.3).

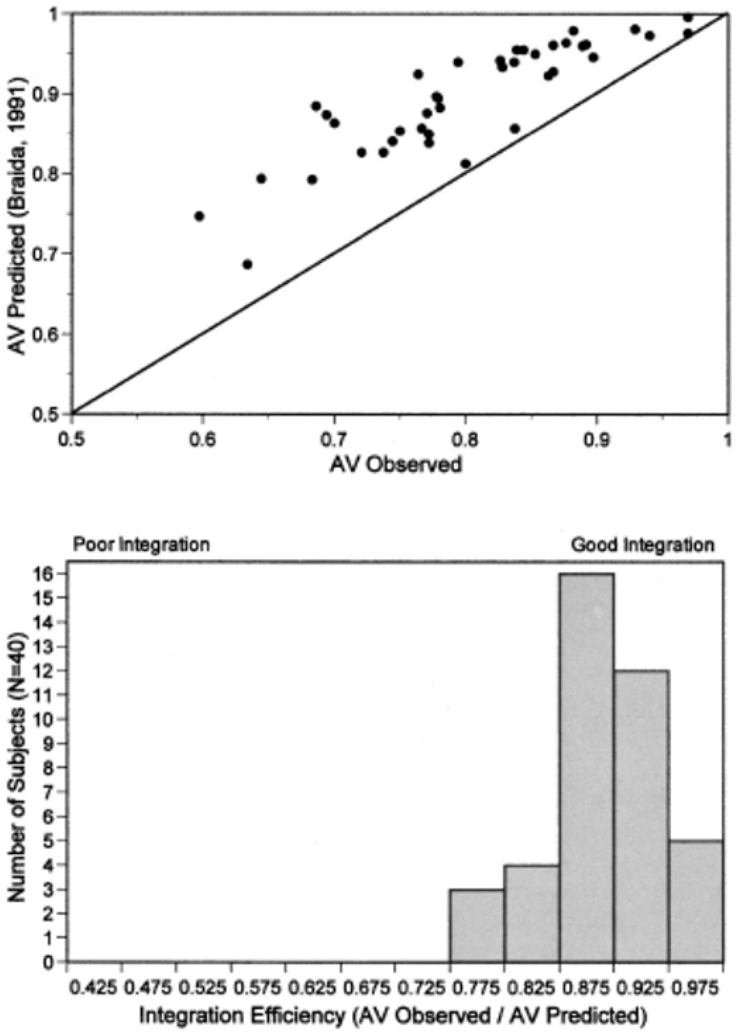

Figure 0.3 PRE model: Observed and derived measures obtained from experiment on consonant recognition in noise (40 subjects). (Top) Observed versus predicted PRE AV scores. The line indicates perfect integration efficiency: $\mathrm{IE}^{PRE} = 1$. Predicted and observed AV scores for several subjects fall near the main diagonal, whereas observed scores for other subjects are significantly less than predicted. (Bottom) Histogram showing distribution of IE^{PRE} values across subjects (from Grant and Seitz, 1998).

0.5.2.3 Integration efficiency based on Fechnerian Scaling

The validity of any IE measure derived from a model of AV speech integration, like the prelabeling model (PRE), depends on the specific assumptions

of the model being valid empirically. We briefly discuss an alternative, less restrictive approach based on a theory of computing *subjective distances* on very general stimulus sets (Dzhafarov and Colonius, 2006).

Recall that a *metric* is a nonnegative function d defined on pairs (x, y) from a set X, say, such that for all $x, y, z \in X$

(i) $d(x, y) \geq 0$ and $d(x, y) = 0$ implies $x = y$;
(ii) $d(x, y) = d(y, x)$;
(iii) $d(x, y) + d(y, z) \geq d(x, z)$.

The theory of *Fechnerian Scaling* (FS) (see, e.g. Dzhafarov and Colonius, 2007) deals with the computation of subjective distances among stimuli from their pairwise discrimination probabilities. The latter are the probabilities with which the judgment "these two stimuli are different" is chosen over "these two stimuli are the same":

$$\psi(x, y) = \mathrm{P}\,[\text{subject judges } x \text{ and } y \text{ in } (x, y) \text{ to be different}] \tag{0.28}$$

For identification tasks, data from confusion matrices are available instead of discrimination probabilities. The cell in a confusion matrix is the probability that stimulus y is identified as stimulus x, denoted as $\eta(x, y)$ for all x, y in the stimulus set X. Thus, we need the additional assumption that

$$1 - \psi(x, y) = \eta(x, y).$$

Given $\eta(x, y)$ for all x, y in the stimulus set X, FS allows one to compute a metric G, say, on X satisfying properties (i) to (iii) above. The only necessary and sufficient empirical condition for the construction is *regular maximality*:

$$\eta(x, x) > \max\{\eta(x, y), \eta(y, x)\}. \tag{0.29}$$

for any $x, y \in X, x \neq y$. In other words, when stimulus x is presented, the probability of identifying x as x should be greater than the probability of identifying x as y, a stimulus different from x. Importantly, $\eta(x, x)$ may vary with x and $\eta(x, y)$ may be different from $\eta(y, x)$.

Let us assume that Fechnerian metrics G_A, G_V, and G_{AV} have been computed from the confusion matrices in the auditory, visual, and audiovisual condition, respectively, for each pair of stimuli $\{i, j\}$. The corresponding metric values $G_A(i, j)$, $G_V(i, j)$, and $G_{AV}(i, j)$ are interpreted as subjective distance between the two stimuli under auditory, visual, and audiovisual presentation, respectively. *A-priori*, these three values are unrelated to each other since they are defined on different stimulus sets. On the other hand, there is a natural one-to-one correspondence across the visual, auditory, and bisensory stimulus sets (*i.e.*, visual stimulus $i \leftrightarrow$ auditory stimulus $i \leftrightarrow$

bisensory stimulus component i). Moreover, given that Fechnerian distances on a given set are unique only up to a similarity transformation, *i.e.*, multiplication with a positive constant, one can standardize each of them such that the maximum distance equals one.[15]

If $G_{AV}(i,j)$ is larger than $G_A(i,j)$ or $G_V(i,j)$, this suggests that adding information from the other modality (V or A) increases the subjective distance between i and j. This increase in subjective distance from the unisensory to the bisensory presentation is proposed as indicator of visual, respectively auditory enhancement, in analogy to VE and VA in Section 0.5.2.1:

$$\begin{aligned} \mathrm{VE}^{FS}(i,j) &= \frac{G_{AV}(i,j) - G_A(i,j)}{1 - G_A(i,j)} \\ &= \frac{G_V(i,j) + G_{sAV}(i,j)}{1 - G_A(i,j)}, \end{aligned} \tag{0.30}$$

and

$$\begin{aligned} \mathrm{VE}^{FS}(i,j) &= \frac{G_{AV}(i,j) - G_V(i,j)}{1 - G_V(i,j)} \\ &= \frac{G_A(i,j) + G_{sAV}(i,j)}{1 - G_V(i,j)}, \end{aligned} \tag{0.31}$$

with

$$G_{sAV}(i,j) = G_{AV}(i,j) - [G_A(i,j) + G_V(i,j)]$$

denoting the superadditivity term, in analogy to Equation (0.24).

In order to derive an overall index of integration efficiency, averaging across all superadditivity terms results in an *Fechnerian Scaling-based multisensory integration efficiency* index:

$$\mathrm{IE}^{FS} = \binom{N}{2}^{-1} \sum_{\{i,j\} \subset S} G_{sAV}(i,j), \tag{0.32}$$

$i \neq j$, with N denoting the number of stimuli in stimulus set S.

The Fechnerian Scaling-based approach to integration efficiency presented here and the prelabeling model (PRE) share the idea of converting the information contained in the confusion matrices into a representation of subjective distances between the stimuli. An important difference is that the

[15] Importantly, Fechnerian distances are always a function of the entire (base) set used to compute them, and the G values are not monotonically related to the probabilities $\eta(x,y)$, although they have been found to correlate highly in many empirical data sets. Moreover, it seems plausible that Fechnerian distances for corresponding stimulus pairs are measured in the same "units".

FS-based approach neither requires explicit assumptions about the space (e.g., Euclidean) and its dimensionality nor any parameter estimation.

One can argue that the definition of IE^{FS} being based on superadditivity, is somewhat arbitrary. Nonetheless, Colonius and Diederich (2007) report on a small data set, a reduced confusion matrix for consonants /b/, /d/, and /g/ presented in Braida et al. (1998). Table 0.2 lists all 3 confusion matrices (auditory, visual, auditory-visual) together with their corresponding Fechnerian distances G_A, G_V, and G_{VA}.

$\psi_A = 1 - \eta_A$ G_A	"b"	"d"	"g"
	0.437	0.717	0.846
b–	0.000	0.450	0.589
	0.700	0.530	0.757
d–	0.450	0.000	0.350
	0.746	0.689	0.566
g–	0.589	0.350	0.000
$\psi_V = 1 - \eta_V$ G_V			
	0.022	0.983	0.996
–b	0.000	1.805	1.527
	0.990	0.146	0.871
–d	1.805	0.000	0.864
	0.989	0.575	0.436
–g	1.527	0.864	0.000
$\psi_{AV} = 1 - \eta_{AV}$ G_{AV}			
	0.007	0.996	0.998
bb	0.000	1.860	1.704
	0.997	0.126	0.876
dd	1.860	0.000	1.203
	0.991	0.731	0.278
gg	1.704	1.203	0.000

Table 0.2 *Each cell: ψ at top and G (Fechnerian distances) at bottom, for auditory (A), visual (V), and audiovisual (AV) presentation (rows $\equiv$ stimuli, columns $\equiv$ responses) with resulting value of* $\mathrm{IE}^{FS} = 0.8737$.

The value of IE^{FS} was computed[16] as 0.8737, which is very close to the correct identification score (87.1 %) predicted by the PRE model (Braida et al., 1998) for the same data set. In general, however, most of the indexes of audiovisual integration efficiency presented here have some degree of arbitrariness and will have to prove their utility and cross-study consistency in future research.

0.6 Measures based on MI modeling of RTs

The focus of this chapter has so far been on measuring MI, rather than modeling. Yet PS, which is a model, has emerged several times as benchmark: any improvement (response speed reduction, improved detection probability, etc.) beyond the level predicted by PS has been defined as measure of MI. In keeping with this approach, we will define CRE as a function of the enhancement observed beyond what is predicted by a particular MI model under consideration. Given that these models typically require estimation of some parameters, the idea here is to estimate them from the unisensory conditions only and subsequently insert these estimates into the MI model in order to predict bisensory RTs. Measures of MI then assess by how much these model predictions fall short of the observed bisensory data. Given the multitude of integration models, however, we need to be selective and will only sketch a few modeling approaches with respect to how they estimate and predict the amount of MI.

0.6.1 Coactivation models

Coactivation is a generic term suggested by Miller (1982) to describe models that allow activation from different channels (in particular, modalities) to combine in satisfying a single criterion for response initiation, in distinction to *separate activation* models (or, race models), where the system never combines activation from different channels in order to meet its criterion for responding (ibid., p. 248). Coactivation models differ with respect to their state space, i.e., whether the state space within which combination is performed, is continuous or discrete. We consider measures of MI for continuous-time models with either discrete and continuous state space. Discrete-time coactivation models are not considered here because of our emphasis on response time measurements.

[16] The IE^{FS} index used was based on the superadditivity term $G_{sAV}(i,j)$ written as ratio rather than difference.

The (Poisson) superposition model. Presentation of a stimulus induces a neural renewal (counting) process[17], $\{N(t), t \geq 0\}$, with interarrival times $\{X_n, n = 1, 2 \dots\}$. Let $W(n) = \sum_{i=1}^{n} X_i$ be the waiting time for the n-th counts. The assumption is that a response is initiated as soon as a fixed number of counts, c, is reached. Note that

$$P(N(t) \geq c) = P(W(c) \leq t).$$

Finally, the observable RT is assumed to be additively composed of the waiting time plus all processes following (or preceding) it. The duration of these additional processes, which may include motor preparation and response execution components, are represented by a random variable M:

$$RT = W(c) + M.$$

The superposition assumption holds that the unisensory renewal processes, $N_V(t)$ and $N_A(t)$, are simply added, defining a new renewal process, so that the waiting time for the c-th count is reduced; specifically, if the visual stimulus is presented τ msec ($\tau > 0$) before the auditory,

$$N_{VA}(t) = N_V(t) + N_A(t - \tau),$$

where $N_A(t - \tau) = 0$ for $t < \tau$.

Under the simplest renewal process (Poisson), expected waiting time for the bisensory condition can be computed as

$$\mathrm{E}W_{V\tau A}(c) = \frac{c}{\alpha_V} - \frac{\alpha_A}{\alpha_V(\alpha_V + \alpha_A)} \exp(-\alpha_V \tau) \sum_{i=0}^{c-1} \frac{(\alpha_V \tau)^i}{i!} (c - i), \quad (0.33)$$

where α_V and α_A are the Poisson intensity parameters for the visual and auditory stimulus, respectively.[18] For $\tau = 0$, this reduces to $c/(\alpha_V + \alpha_A)$.

Let $\mathrm{E}RT_{V\tau A} = \mathrm{E}W_{V\tau A}(c) + \mathrm{E}M$. In obvious notation, we define as measure of crossmodal response enhancement for the Poisson superposition model,

$$\mathrm{CRE}_{SUP,\tau} = \frac{\mathrm{E}RT_{V\tau A} - \mathrm{E}RT^{obs}_{V\tau A}}{\mathrm{E}RT_{V\tau A}} \times 100, \quad (0.34)$$

assuming parameters c and $\mathrm{E}M$ to be invariant across the unisensory and bisensory conditions. Note that $\mathrm{CRE}_{SUP,\tau}$ increases as a function of c; thus, the Poisson superposition model is consistent with the prediction of inverse effectiveness. On the other hand, it cannot predict inhibition.

[17] For exact definitions, see e.g. Ross (1996).
[18] For $\tau < 0$, τ must be replaced by $-\tau$ and α_V and α_A interchanged.

Diffusion models. In these models, presentation of a stimulus is assumed to induce a stochastic process that is often described by a linear, first-order stochastic differential equation[19] of the form

$$dX(t) = \mu(X(t), t) + \sigma(X(t), t)\, dW(t), \tag{0.35}$$

where $W(t)$ is a standard *Wiener process*, $\mu(x, t)$ is called the *effective drift rate* describing the instantaneous rate of expected increment change at time t and state $x = X(t)$. Factor $\sigma(X(t), t)$ in front of the instantaneous increments $dW(t)$ is called *diffusion coefficient* relating to the variance of the increments.

Modeling information accumulation and predicting response times, however, requires one to make concrete assumptions on drift rates and diffusion coefficients resulting in a large variety of stochastic diffusion models. For example, setting $\mu(x, t) = \delta$ and $\sigma(X(t), t) = \sigma$ defines a time-homogeneous Wiener process with drift (setting $\delta = 0$ is the standard Wiener process). The drift rate is interpreted as describing the rate of information accumulation under different stimulus conditions.

Termination of the accumulation process is then defined by the first time it reaches a threshold, C ($C > 0$). This *stopping time*, denoted as ν, is the smallest value for t such that $X(t) = C$. If $X(0) = 0$, expected stopping time in the Wiener process with drift δ is

$$\mathrm{E}[\nu \mid X(0) = 0] = C/\delta, \tag{0.36}$$

which independent of the diffusion coefficient. Observed RT is defined as the sum of (random variables) ν and a non-decision component M, $RT = \nu + M$.

Applying this model version to the redundant signals paradigm, we assume two Wiener processes with drift rates δ_V and δ_A, respectively, for the unisensory conditions. In the bisensory condition with $SOA \equiv \tau = 0$, a superposed Wiener process is defined by

$$X_{VA}(t) = X_V(t) + X_A(t) \tag{0.37}$$

with drift rate $\delta_V + \delta_A$, while postulating identical threshold values C and mean values of M, for all conditions. Given the expected stopping times $C/\delta_V, C/\delta_A, C/(\delta_V + \delta_A)$, one can define a measure of crossmodal enhancement exactly like Equation (0.34) for the Poisson superposition model at $\tau = 0$. Obviously, however, under these simplified assumptions the two models become indistinguishable, predicting the same amount of enhancement. The problem dissolves when predictions for non-simultaneous stimuli for two

[19] For exact definitions, we must refer to the literature.

modalities (Schwarz, 1994) or more (Diederich, 1995) are derived and CRE measures analogous to Equation (0.34) can be defined:

$$\mathrm{CRE}_{DIF,\tau} = \frac{\mathrm{E}RT_{V\tau A} - \mathrm{E}RT^{obs}_{V\tau A}}{\mathrm{E}RT_{V\tau A}} \times 100. \tag{0.38}$$

Moreover, for $\tau = 0$, setting $\mu(x,t) = \delta - \gamma x$ and $\sigma(x,t) = \sigma$ defines a time-homogeneous *Ornstein-Uhlenbeck process* (OUP)[20] . For $\gamma > 0$ this implies that the accumulation rate decays in dependence of the current state x (e.g., Diederich, 1995). Given that for this and related models, expected stopping times are often not available in closed form, crossmodal enhancement measures of the form of Equation (0.38) may be approximated by simulation or, alternatively, by Markov chain approximation.[21]

0.6.2 Time-window-of-integration framework

While the PS mechanism by itself constitutes a broad class of models at both the neural and behavioral level, simple race models often do not fare too well empirically and, as mentioned, typically only serve as point of reference in defining an enhancement measure (see Section 0.4.3). The time-window-of-integration (TWIN) framework for response speed, measured as manual or saccadic RT, is a simple extension of the PS model. The amount of RT facilitation not accounted for by the latter (compare Equation 0.17) equals,

$$\mathrm{E}\min\{RT_V, RT_A\} - \mathrm{E}RT_{VA}$$

where RT_A and RT_V are the observed latencies of unisensory responses. Here, $\mathrm{E}\min\{RT_V, RT_A\}$ refers to the RT predicted by a PS rule (stochastically independent or dependent race) and $\mathrm{E}RT_{VA}$ is the observed bisensory mean RT. The TWIN framework postulates two serial processing stages. A first (race) stage among the activity elicited by the different modalities is followed by a second stage that is defined by default: it includes all subsequent, possibly temporally overlapping, processes that are not part of the processes in the first stage, and crossmodal interaction can only occur in the second stage.

While the framework is mute about the specific mechanism of integration in the second stage, its central feature is the notion of a time-window of MI. It postulates that crossmodal interaction occurs *only* if the peripheral processes of the first stage all terminate within a given temporal interval, the 'time

[20] But $\tau \neq 0$ implies a non-time-homogenous OU process.

[21] Roughly, after fitting the unisensory data with an OUP model each, sample unisensory values $x_V(t)$ and $x_A(t)$ for any t, add them to define a superposed process, and estimate the expected stopping time of that process.

window of integration'. The result of crossmodal interaction manifests itself in an increase, or decrease, of second stage processing time. The window acts as a filter determining whether afferent information delivered from different sensory organs is registered close enough in time to trigger MI. 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.[22] Although the amount of interaction does not directly depend on stimulus onset asynchrony (SOA) of the stimuli, temporal tuning of the interaction yet occurs because the *probability of the integration event* is modulated by the SOA value. Formalization of the framework makes these observations explicit.

We introduce some notation and derive an expression for the measure of MI in the TWIN framework, With τ ($-\infty < \tau < +\infty$) as SOA value and ω ($\omega \geq 0$) as parameter for the integration window width, these assumptions imply that the event that MI occurs, denoted by I, equals

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

where T_V, T_A are assumed to be continuous random variables and the presentation of the visual stimulus is (arbitrarily) defined as the physical zero time point. Thus, the probability of integration to occur, $P(I)$, is an increasing function of ω, but its dependence on τ will be a function of the specific distributions assumed for T_V and T_A.

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

$$\begin{aligned} \mathrm{E}[RT_{V\tau A}] &= \mathrm{E}[S_1] + P(I)\,\mathrm{E}[S_2|I] + [1 - P(I)]\,\mathrm{E}[S_2|I^c] \\ &= \mathrm{E}[S_1] + \mathrm{E}[S_2|I^c] - P(I) \times \Delta. \\ &= \mathrm{E}[\min(T_V, T_A + \tau)] + \mathrm{E}[S_2|I^c] - P(I) \times \Delta. \end{aligned} \tag{0.39}$$

Here, I^c denotes the event complementary to I and Δ stands for $\mathrm{E}[S_2|I^c] - \mathrm{E}[S_2|I]$. 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. Because event I cannot occur in the unimodal (visual or auditory) condition, expected RT under these

[22] Note that the window of the TWIN framework is only defined temporally, in contrast to a spatio-temporal window sometimes postulated.

conditions is, respectively,

$$\mathrm{E}[RT_V] = \mathrm{E}[T_V] + \mathrm{E}[S_2|I^c] \quad \text{and} \quad \mathrm{E}[RT_A] = \mathrm{E}[T_A] + \mathrm{E}[S_2|I^c].$$

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

$$SFE \equiv \min\{\mathrm{E}[T_V], \mathrm{E}[T_A] + \tau\} - \mathrm{E}[\min\{T_V, T_A + \tau\}].$$

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

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

Thus, in the TWIN framework crossmodal *facilitation* observed in a redundant signals task may be due to MI or statistical facilitation, or both. This shows that the TWIN extends the race model class by predicting integration effects over and above statistical facilitation. Moreover, a potential multisensory *inhibitory* effect occurring in the second stage may be weakened, or even masked completely, by the simultaneous presence of statistical facilitation in the first stage.

We have shown that one can derive various empirically testable predictions from the TWIN framework even without assuming specific distributions for the random processing times. In addition, when T_V and T_A are independent and exponentially distributed random variables and the expected value for 2nd-stage processing time with no crossmodal interaction is set as parameter μ, then numerical estimates of the overall crossmodal interaction effect, $SFE + P(I) \times \Delta$, are available. This suggests the following definition for crossmodal enhancement:

$$\mathrm{CRE}_{TWIN} = \frac{\mathrm{E}RT_{V\tau A} - \mathrm{E}RT^{obs}_{V\tau A}}{\mathrm{E}RT_{V\tau A}} \times 100, \tag{0.40}$$

with $\mathrm{E}RT^{obs}_{V\tau A}$ denoting observed mean bisensory RT and $\mathrm{E}RT_{V\tau A}$ the expected bisensory RT under the TWIN model, which can be calculated using parameter estimates obtained from fitting the model to the observations.

Note that "temporal window of integration" has become an important concept in describing crossmodal binding effects as function, e.g., of age, specific disorders, or training in a variety of MI tasks apart from RTs[23]. In fact, the width of the time window can by itself be taken as measure for MI:

[23] It is worth pointing out that the time window concept in the TWIN framework differs from the one used in most empirical studies. The latter is typically defined by the range of SOA values wherein crossmodal effects can be observed. In contrast, in the former (i) window width is a parameter to be estimated from the data, and (ii) the filter is not in principle limited to the temporal structure of the stimulus context but could be defined more broadly, e.g. including spatial features or subjective values see, e.g., (Bean et al., 2021).

In the temporal order judgment (TOJ) task, where subjects are required to judge the order of stimuli (visual first vs. auditory first), 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. Within a simple extension of the TWIN framework to include the TOJ task, 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 (Diederich and Colonius, 2015).

0.7 Conclusions

It turned out that, in order to construct valid measures of integration, a possible effect of PS had to be taken into account, in both behavioral and neural contexts. Specifically, we have argued that the common index for RTs in the redundant signals paradigm (see Equation 0.2),

$$\mathrm{CRE}_{RT} = \frac{\min\{\mathrm{E}RT_V, \mathrm{E}RT_A\} - \mathrm{E}RT_{VA}}{\min\{\mathrm{E}RT_V, \mathrm{E}RT_A\}} \times 100, \quad (0.2)$$

should be replaced by assuming a race model with maximal negative dependence

$$\mathrm{CRE}_{RT}^{min} = \frac{\mathrm{E}^{(-)} \min\{RT_V, RT_A\} - \mathrm{E}RT_{VA}}{\mathrm{E}^{(-)} \min\{RT_V, RT_A\}} \times 100,$$

which is Equation 0.17 for $\tau = 0$. The latter is a more conservative index because it allows for the possibility that the "race" between visual and auditory activation may be (maximally) negatively dependent in the statistical sense, that is, it measures how much faster observed mean RT is than the fastest one that can be generated by PS alone.

A further argument in favor of using $\mathrm{CRE}_{RT}^{(-)}$ is that $\mathrm{E}^{(-)} \min\{RT_V, RT_A\}$ can be sensitive to the shape of the entire distribution of the unisensory RT distributions, like moments higher than the mean, see Colonius and Diederich (2017). Another, non-RT, example is a discrimination task where estimator variance is required to obtain a statistically optimal linear combination of modalities (Ernst and Banks, 2002; Drugowitsch et al., 2014), so that any MI measure gauging the degree of deviation from optimality will be a function of the second moment.

Thus, instead of defining MI measures via means only, it may be argued

Type	Index	Definition	Section
spikes	CRE_{SP}	$\frac{\mathrm{E}N_{VA}-\max\{\mathrm{E}N_V,\mathrm{E}N_A\}}{\max\{\mathrm{E}N_V,\mathrm{E}N_A\}} \times 100,$	0.3.3
	CRE^{*}_{SP}	$\frac{\mathrm{E}N_{VA}-\mathrm{E}\max\{N_V,N_A\}}{\mathrm{E}\max\{N_V,N_A\}} \times 100$	0.3.3
	CRE^{max}_{SP}	$\frac{\mathrm{E}N_{VA}-\mathrm{E}^{(-)}\max\{N_V,N_A\}}{\mathrm{E}^{(-)}\max\{N_V,N_A\}} \times 100$	0.3.3
RTs	$\mathrm{CRE}_{RT,\tau}$	$\frac{\min\{\mathrm{E}RT_V,\mathrm{E}RT_A+\tau\}-\mathrm{E}RT_{V\tau A}}{\min\{\mathrm{E}RT_V,\mathrm{E}RT_A+\tau\}} \times 100$	0.4.1
	$\mathrm{CRE}_{RT,\tau_1\tau_2}$	$\frac{\min\{\mathrm{E}RT_V,\mathrm{E}RT_A+\tau_1,\mathrm{E}RT_T+\tau_1+\tau_2\}-\mathrm{E}RT_{V\tau_1A\tau_2T}}{\min\{\mathrm{E}RT_V,\mathrm{E}RT_A+\tau_1,\mathrm{E}RT_T+\tau_1+\tau_2\}} \times 100$	0.4.1
	$\mathrm{CRE}_{RT,(\tau_1)\tau_2}$	$\frac{\mathrm{E}RT_{V\tau_1A}-\mathrm{E}RT_{V\tau_1A\tau_2T}}{\mathrm{E}RT_{V\tau_1A}} \times 100$	0.4.1
	CRE^{min}_{RT}	$\frac{\mathrm{E}^{(-)}\min\{RT_V,RT_A\}-\mathrm{E}RT_{VA}}{\mathrm{E}^{(-)}\min\{RT_V,RT_A\}} \times 100$	0.4.1
accuracy	CRE_{DR}	$\frac{p_{VA}-\max\{p_V,p_A\}}{\max\{p_V,p_A\}} \times 100$	0.5.1
	CRE_{SDT}	$\frac{d'_{VA}-\max\{d'_V,d'_A\}}{\max\{d'_V,d'_A\}} \times 100$	0.5.1
	CRE^{PS}_{SDT}	$\frac{d'_{VA}-d'^{PS}_{VA}}{d'^{PS}_{VA}} \times 100$	0.5.1
AV speech	p_{sAV} (superadditivity)	$p_{AV}-(p_A+p_V)$	0.5.2
	VE (vis. enhancement)	$\frac{p_{AV}-p_A}{1-p_A} = \frac{p_V+p_{sAV}}{1-p_A}$	0.5.2
	AE (aud. enhancement)	$\frac{p_{AV}-p_V}{1-p_V} = \frac{p_A+p_{sAV}}{1-p_V}$	0.5.2
	p^{I}_{AV}	$1-(1-p_A)\times(1-p_V)$	0.5.2
	IE^{I} (integr. efficiency)	$\frac{p^{obs}_{AV}-p^{I}_{AV}}{1-p^{I}_{AV}}$	0.5.2
	IE^{PRE}	$d'_{AV}(obs)/d'_{AV}(pred)$	0.5.2
	$G_{sAV}(i,j)$	$G_{AV}(i,j)-[G_A(i,j)+G_V(i,j)]$	0.5.2
	IE^{FS}	$\binom{N}{2}^{-1}\sum_{\{i,j\}\subset S} G_{sAV}(i,j)$	0.5.2
RT model	$\mathrm{CRE}_{SUP,\tau}$	$\frac{\mathrm{E}RT_{V\tau A}-\mathrm{E}RT^{obs}_{V\tau A}}{\mathrm{E}RT_{V\tau A}} \times 100$	0.6.1
	$\mathrm{CRE}_{DIF,\tau}$	$\frac{\mathrm{E}RT_{V\tau A}-\mathrm{E}RT^{obs}_{V\tau A}}{\mathrm{E}RT_{V\tau A}} \times 100$	0.6.1
	CRE_{TWIN}	$\frac{\mathrm{E}RT_{V\tau A}-\mathrm{E}RT^{obs}_{V\tau A}}{\mathrm{E}RT_{V\tau A}} \times 100$	0.6.2

Table 0.3 *List of all indexes in the chapter (for spikes, RTs, detection accuracy, AV speech identification, RT models)*

that one should compare entire distributions in order to obtain more informative measures. Assume there exists a numerical function δ measuring the distance between two distributions, e.g. $\delta(F_A, F_{VA})$, one may define crossmodal response enhancement, in analogy to CRE_{RT} above, by

$$\mathrm{CRE}_\delta = \min\{\delta(F_V, F_{VA}), \delta(F_A, F_{VA})\} \times 100. \tag{0.41}$$

Here, δ is already normalized to a range from zero to one; if F_{VA} is equal to one of the unisensory distributions, then $\mathrm{CRE}_\delta = 0$.[24] Thus, the first two requirements for a CRE measure (see Section 0.2.2) are satisfied, while the inhibition case is not covered. More complex measures are certainly possible; however, a more pressing task is to find criteria for selecting some measure δ from the "universe" of distance measures between distributions that would make the choice less arbitrary.

0.8 Bibliographical notes

Despite the limited scope of this chapter, we hope to have given a first glimpse into the various ways and issues of defining measures of MI. A broader and deeper view may be gained from the references given in this section.

A number of comprehensive handbooks and review articles on MI are available: Calvert et al. (2004); Naumer and Kayser (2010); Stein (2012); Bremner et al. (2012); Murray and Wallace (2012); Stevenson et al. (2014); van Opstal (2016); Colonius and Diederich (2020). The first monograph on MI from the neurophysiology point of view is Stein and Meredith (1993), while Stein et al. (2009) discuss quantitative methods for measuring MI at the single-neuron level. Early studies by Todd (1912), measuring RT to stimuli from two or more sensory modalities, presented both singly and together, are often seen as the beginnings of the scientific study of crossmodal behavior. Raab (1962) is the classic reference for a treatment of the "race model" and PS mechanisms for RTs. The latter has typically been presented under the hypothesis of stochastic independence. The "race model inequality" (see Equation 0.15), first developed in Miller (1982) and tested in Diederich and Colonius (1987), initiated the discussion of non-independent PS in the context of copula theory (Colonius, 1990, 2016; Colonius and Diederich, 2017) and the development of related statistical tests (Ulrich et al., 2007; Gondan, 2010; Gondan et al., 2012; Lombardi et al., 2019). Generalized race

[24] Probability summation could be accounted for by defining $\mathrm{CRE}_\delta = \delta(\min\{F_V(t) + F_A(t), 1\}, F_{VA}) \times 100$.

model inequalities have been discussed, e.g., in Colonius et al. (2017); Gondan et al. (2020); Gondan and Vorberg (2021). The "principle of congruent effectiveness" (Otto et al., 2013), stating that multisensory behavior (specifically, speedup of response times) is largest when behavioral performance in corresponding unisensory conditions is similar, corresponds to the index of unisensory imbalance (UI) (see Equation 0.3).

Regarding accuracy measures, Jones (2016) provides a comprehensive tutorial about models of cue combination based on measures of sensitivity including signal detection theory (Macmillan and Creelman, 2005; Wickens, 2002). Schwarz and Miller (2014) point out that PS does not always lead to facilitation in compound detection and discrimination tasks because an increase of hit rate may also cause an increase of false alarms; evaluating uni-vs. bisensory performance should, therefore, be performed via comparing the associated areas under the ROC curves. Billock et al. (2021) present a framework for comparing spike rates from AV integration in cortical bisensory neurons with psychophysical (discrimination) data and suggest vector-like Minkowski combination models describing either.

The literature on AV speech processing is huge, the handbook by Bailly et al. (2012) is a good source, as well as reports from the *International Conference on Auditory-Visual Speech Processing* (AVSP).[25] More details on the Fechnerian Scaling approach is found in the chapter by E.N. Dzhafarov and H. Colonius in this volume (*Fechnerian Scaling: Dissimilarity Cumulation Theory*).

The Poisson superposition model for MI has been introduced in Schwarz (1989) and discussed in Diederich and Colonius (1991); Diederich (1995), and Schwarz (1994). A tutorial on diffusion processes for RTs is given in Smith (2000), and a comprehensive treatment of stochastic models for decision-making is the chapter by Diederich & and Mallahi-Karai in Volume II (Diederich and Mallahi-Karai, 2018). Notably, diffusion models can be extended to describe binary choice response tasks by assuming an upper and a lower absorbing bound for the accumulation process (Ratcliff, 1978). Such a diffusion superposition model for audiovisual data is discussed and tested by experiment in Blurton et al. (2014). Drugowitsch et al. (2014) introduce a diffusion model for visual-vestibular integration with a weighted superposition approach that accumulates evidence optimally across both cues and time. For other extensions of diffusion models see e.g. Mallahi-Karai and Diederich (2021); Diederich and Oswald (2016); Diederich (1997). The time-window-of-integration model was introduced by the authors in 2004 (Colo-

[25] http://www.isca-speech.org/archive

nius and Diederich, 2004) and subsequently extended and experimentally tested in **?**Diederich and Colonius (2007, 2008a,b).

References

Bailly, G., Perrier, P., and Vatikiotis-Bateson, E. (eds). 2012. *Audiovisual speech processing*. Cambridge University Press.

Bean, N.L., Stein, B.E., and Rowland, B.A. 2021. Stimulus value gates multisensory integration. *European Journal of Neuroscience*, **53**, 3142–3159.

Billock, V.A., Kinney, M.J., Schnupp, J.W.H., and Meredith, M.A. 2021. A simple vector-like law for perceptual information combination is also followed by a class of cortical multisensory bimodal neurons. *iScience*, https://doi.org/10.1016/j.isci.2021.102527.

Blurton, S.P., Greenlee, M.W., and Gondan, M. 2014. Multisensory processing of redundant information in go/no–go and choice responses. *Attention, Perception, and Psychophysics*, **76**, 1212–1233.

Braida, L.D. 1991. Crossmodal integration in the identification of consonant segments. *Quarterly Journal of Experimental Psychology A*, **43**(3), 647–677.

Braida, L.D., Sekiyama, K., and Dix, A.K. 1998. Integration of audiovisually compatible and incompatible consonants in identification experiments. Pages 49–54 of: Burnham, D., Robert-Ribes, J., and Vatikiotis-Bateson, E. (eds), *Auditory-Visual Speech Processing 1998 (AVSP98)*. International Conference on Auditory-Visual Speech Processing 1998, Terrigal, Sydney, NSW, Australia.

Bremner, A.J., Lewkowicz, D.J., and Spence, C. (eds). 2012. *Multisensory Development*. Oxford, UK: Oxford University Press.

Calvert, G., Spence, C., and Stein, B. E. 2004. *Handbook of multisensory processes*. MIT Press.

Colonius, H. 1990. Possibly dependent probability summation of reaction time. *Journal of Mathematical Psychology*, **34**, 253–275.

Colonius, H. 2016. 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**, 2–10.

Colonius, H., and Diederich, A. 2004. Multisensory interaction in saccadic reaction time: a time-window-of-integration model. *Journal of Cognitive Neuroscience*, **16**, 1000–1009.

Colonius, H., and Diederich, A. 2007. A measure of auditory-visual integration efficiency based on Fechnerian Scaling. In: Vroomen, J., Swerts, M., and Krahmer, E. (eds), *Auditory-Visual Speech Processing 2007 (AVSP2007)*. International Conference on Auditory-Visual Speech Processing 2007, Kasteel Groenendaal, Hilvarenbeek, The Netherlands, no. Paper 33.

Colonius, H., and Diederich, A. 2017. Measuring multisensory integration: from reaction times to spike counts. *Scientific Reports*, **7**(1), 3023. http://dx.doi.org/10.1038/s41598-017-03219-5.

Colonius, H., and Diederich, A. 2020. Formal models and quantitative measures of multisensory integration: a selective overview. *European Journal of Neuroscience*, **51**, 1161–1178.

Colonius, H., Diederich, A., and Steenken, R. 2009. Time-window-of-integration (TWIN) model for saccadic reaction time: effect of auditory masker level on visual-auditory spatial interaction in elevation. *Brain Topography*, **21**, 177–184.

Colonius, H., Woff, F.H., and Diederich, A. 2017. Trimodal race model inequalities in multisensory integration. *Frontiers in Psychology*, **8**(1141).

Dias, J.W., McClaskey, C.M., and Harris, K.C. 2021. Audiovisual Speech Is More Than the Sum of Its Parts: Auditory-Visual Superadditivity compensates for age-related declines in audible and lipread speech intelligibility. *Psychology and Aging*, **36**(4), 520–530.

Diederich, A. 1995. Intersensory facilitation of reaction time: evaluation of counter and diffusion coactivation models. *Journal of Mathematical Psychology*, **39**, 197–215.

Diederich, A. 1997. Dynamic stochastic models for decision making under time constraints. *Journal of Mathematical Psychology*, **41**, 260–274.

Diederich, A., and Colonius, H. 1987. Intersensory facilitation in the motor component? *Psychological Research*, **49**, 23–29.

Diederich, A., and Colonius, H. 1991. A further test of the superposition model for the redundant signals effect in bimodal detection. *Perception & Psychophysics*, **50**, 83–83.

Diederich, A., and Colonius, H. 2007. Modeling spatial effects in visual-tactile reaction time. *Perception and Psychophysics*, **69**(1), 56–67.

Diederich, A., and Colonius, H. 2008a. Crossmodal interaction in saccadic reaction time: Separating multisensory from warning effects in the time window of integration model. *Experimental Brain Research*, **186**(1), 1–22.

Diederich, A., and Colonius, H. 2008b. When a high-intensity "distractor" is better then a low-intensity one: Modeling the effect of an auditory or tactile nontarget stimulus on visual saccadic reaction time. *Brain Research*, **1242**, 219–230.

Diederich, A., and Colonius, H. 2015. The time window of multisensory integration: Relating reaction times and judgments of temporal order. *Psychological Review*, **122**(2), 232–241.

Diederich, A., and Mallahi-Karai. 2018. Stochastic methods for modeling decision-making. Chap. 1, pages 1–70 of: Batcheldeer, W.H., Colonius, H., and Dzhafarov, E.N. (eds), *New handbook of mathematical psychology*, vol. II. Cambridge University Press.

Diederich, A., and Oswald, P. 2016. Multi-stage sequential sampling models with finite or infinite time horizon and variable boundaries. *Journal of Mathematical Psychology*, **74**, 128–145.

Drugowitsch, J., DeAngelis, G.C., Klier, E. M., Angelaki, D.E., and Pouget, A. 2014. Optimal multisensory decision-making in a reaction-time task. *eLife*, e03005. doi: 10.7554/eLife.03005.

Dzhafarov, E.N., and Colonius, H. 2006. Reconstructing distances among objects from their discriminability. *Psychometrika*, **71**(2), 365–386.

Dzhafarov, E.N., and Colonius, H. 2007. Dissimilarity cumulation theory and subjective metrics. *Journal of Mathematical Psychology*, **51**, 290–304.

Ernst, M., and Banks, M.S. 2002. Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, **415**, 429–433.

Fréchet, M. 1951. Sur les tableaux de corrélation dont les marges sont donnés. *Annales de l'Université de Lyon*, **Section A, Séries 3**(14), 53–77.

Gondan, M. 2010. A permutation test for the race model inequality. *Behavior Research Methods*, **42**, 23–28.

Gondan, M., and Vorberg, D. 2021. Testing trisensory interactions. *Journal of Mathematical Psychology*, **101**, https://doi.org/10.1016/j.jmp.2021.102513.

Gondan, M., Riehl, V., and Blurton, S.P. 2012. Showing that the race model inequality is not violated. *Behavior Research Methods*, **44**, 248–255.

Gondan, M., Dupont, D., and Blurton, S.P. 2020. Testing the race model in a difficult redundant signals task. *Journal of Mathematical Psychology*, **95**, https://doi.org/10.1016/j.jmp.2020.102323.

Grant, K.W. 2002. Measures of auditory-visual integration for speech understanding: A theoretical perspective (L). *Journal of the Acoustical Society of America*, **112**(1), 30–33.

Grant, K.W., and Seitz, P.F. 1998. Measures of auditory–visual integration in nonsense syllables and sentences. *Journal of the Acoustical Society of America*, **104**(4), 2438–2450.

Green, D.M., and Swets, J.A. 1974. *Signal detection theory and psychophysics*. New York, NY: Robert E. Krieger Publishing Co.

Joe, H. 1997. *Multivariate models and dependence concepts*. Monographs on Statistics and Applied Probability, no. 73. London, UK: Chapman & Hall.

Jones, P.R. 2016. A tutorial on cue combination and Signal Detection Theory: Using changes in sensitivity to evaluate how observers integrate sensory information. *Journal of Mathematical Psychology*, **73**, 117–139.

Lombardi, L., D'Allesandro, M., and Colonius, H. 2019. A new nonparametric test for the racemodel inequality. *Behavior Research Methods*, **51**, 2290–2301.

Lovelace, C.T., Stein, B.E., and Wallace, M.T. 2003. An irrelevant light enhances auditory detection in humans: a psychophysical analysis of multisensory integration in stimulus detection. *Cognitive Brain Research*, **17**, 447–453.

Macmillan, N.A., and Creelman, C.D. 2005. *Detection theory: a user's guide*. Lawrence Erlbaum Associates.

Mallahi-Karai, K., and Diederich, A. 2021. Decision with multiple alternatives: Geometric models in higher dimensions–the disk model. *Journal of Mathematical Psychology*, **100**(https://doi.org/10.1016/j.jmp.2020.102493).

Massaro, D.W., and Cohen, M.M. 2000. Tests of auditory-visual integration eciency within the framework of the fuzzy logical model of perception. *Journal of the Acoustical Society of America*, **108**(2), 784–789.

Meredith, M.A., and Stein, B.E. 1983. Interactions among converging sensory inputs in the superior colliculus. *Science*, **221**, 389–391.

Miller, J.O. 1982. Divided attention: evidence for coactivation with redundant signals. *Cognitive Psychology*, **14247–279**.

Miller, R.L., Pluta, S.R., Stein, B.E., and Rowland, B.A. 2015. Relative unisensory strength and timing predict their multisensory product. *The Journal of Neuroscience*, **35**(13), 5213–5220.

Murray, M.M., and Wallace, M.T. (eds). 2012. *The Neural Bases of Multisensory Processes*. Frontiers in Neuroscience. Boca Rato, FL: CRC Press.

Naumer, M.J., and Kayser, J. (eds). 2010. *Multisensory Object Perception in the Primate Brain*. New York, NY: Springer-Verlag.

Otto, T.U., Dassy, B., and Mamassian, P. 2013. Principles of multisensory behavior. *The Journal of Neuroscience*, **33**, 7463–7474.

Raab, D.H. 1962. Statistical facilitation of simple reaction time. *Transactions of the New York Academy of Sciences*, **24**, 574–590.

Ratcliff, R. 1978. Theory of memory retrieval. *Psychological Review*, **85**, 59–108.

Ross, S.M. 1996. *Stochastic processes*. Second edn. New York, NY: John Wiley & Sons.

Schwarz, W. 1989. A new model to explain the redundant-signals effect. *Perception &Psychophysics*, **46**(5), 490–500.

Schwarz, W. 1994. Diffusion, superposition, and the redundant-targets effect. *Journal of Mathematical Psychology*, **38**, 504–520.

Schwarz, W., and Miller, J.O. 2014. When less equals more: probability summation without sensitivity improvement. *Journal of Experimental Psychology: Human Perception and Performance*, **40**(5), 2091–2100.

Smith, P.L. 2000. Stochastic dynamic models of response time and accuracy: a foundational primer. *Journal of Mathematical Psychology*, **44**, 408–463.

Stein, B. E. (ed). 2012. *The New Handbook of Multisensory Processes*. MIT Press.

Stein, B. E., Burr, D., Constantinidis, C., Laurienti, P., Meredith, M., Perrault Jr, T., Ramachandran, R., Roeder, B., Rowland, B., Sathian, K., Schroeder, C., Shams, L., Stanford, T., Wallace, M., Yu, L., and Lewkowicz, D. 2010. Semantic confusion regarding the development of multisensory integration: a practical solution. *European Journal of Neuroscience*, **31**, 1713–1720.

Stein, B.E., and Meredith, M.A. 1993. *The merging of the senses*. MIT Press.

Stein, B.E., Stanford, T.R., Ramachandran, R., Perrault Jr, T.J., and Rowland, B.A. 2009. Challenges in quantifying multisensory integration: alternative criteria, models, and inverse effectiveness. *Experimental Brain Research*, **198**, 113–126.

Stevenson, R.A., Ghose, D., Krueger Fister, J., Sarko, D.K., Altieri, N.A., Nidiffer, A.R., Kurela, L.R., Siemann, J.K., James, T.W., and Wallace, M.T. 2014. Identifying and quantifying multisensory integration: a tutorial review. *Brain Topography*, **27**(6), 707–730.

Todd, J.W. 1912. Reaction to multiple stimuli. *Archives of Psychology No. 25. Columbia Contributions to Philosophy and Psychology*, **XXI**(8). New York: The Science Press.

Tye-Murray, N., Sommers, M.S., and Spehar, B. 2007. Audiovisual integration and lipreading abilities of older adults with normal and impaired hearing. *Ear and Hearing*, **28**(5), 656–668.

Tye-Murray, N., Sommers, M., Spehar, B., Myerson, J., and Hale, S. 2010. Aging, audiovisual integration, and the principle of inverse effectiveness. *Ear and Hearing*, **31**, 636–644.

Tye-Murray, N., Spehar, B., Myerson, J., Hale, S., and Sommers, M. 2016. Lipreading and audiovisual speech recognition across the adult lifespan: implications for audiovisual integration. *Psychology and Aging*, **31**(4), 380–389.

Ulrich, R., Miller, J.O., and Schröter, H. 2007. Testing the race model inequality: an algorithm and computer programs. *Behavior Research Methods*, **39**(2), 291–302.

van de Rijt, L.P.H., Roye, A., Mylanus, E.A.M., van Opstal, A.J., and van Wanrooij, M.M. 2019. The principle of inverse effectiveness in audiovisual speech perception. *Frontiers in Human Neuroscience*, **13**(335).

van Opstal, A.J. 2016. *The Auditory System and Human Sound-Localization Behavior*. San Diego: Academic Press. Chap. 13 Multisensory Integration, pages 361–392.

Welch, R.B., and Warren, D.H. 1986. Intersensory interactions. Chap. 25, pages 1–36 of: Boff, K.R., Kaufman, L., and Thomas, J.P. (eds), *Handbook of perception and human performance*, vol. 1. New York, NY: John Wiley & Sons.

Wickens, T.D. 2002. *Elementary signal detection theory*. Oxford University Press.

Yu, L., Cuppini, C., Xu, J., Rowland, B.A., and Stein, B.E. 2019. Cross-modal competition: the default computation for multisensory processing. *The Journal of Neuroscience*, **39**(8), 1374–1385.

Index