Multisensory integration and exogenous spatial attention: a time-window-of-integration analysis

Adele Diederich

Health: Life Sciences & Chemistry

Jacobs University Bremen

Hans Colonius

Department of Psychology

Carl von Ossietzky Universität, Oldenburg

Final Version

Abstract

While it is well-documented that occurrence of an irrelevant and non-predictive sound facilitates motor responses to a subsequent target light appearing nearby, the cause of this ‘exogenous spatial cuing effect’ has been under discussion. On the one hand, it has been postulated to be the result of a shift of visual spatial attention possibly triggered by parietal and/or cortical supramodal ‘attention’ structures. On the other hand, the effect has been considered to be due to multisensory integration based on the activation of multisensory convergence structures in the brain. Recent reaction time experiments have suggested that multisensory integration and exogenous spatial cuing differ in their temporal profiles of facilitation: when the non-target occurs 100 to 200 ms before the target, facilitation is likely driven by crossmodal exogenous spatial attention, whereas multisensory integration effects are still seen when target and non-target are presented nearly simultaneously. Here we develop an extension of the time-window-of-integration (TWIN) model that combines both mechanism within the same formal framework. The model is illustrated by fitting it to data from a focused attention task with a visual target and an auditory non-target presented at horizontally or vertically varying positions. Results show that both spatial cuing and multisensory integration may co-exist in a single trial in bringing about the crossmodal facilitation of reaction time effects. Moreover, the formal analysis via TWIN allows to predict and quantify the contribution of either mechanism as they occur across different spatio-temporal conditions.

Key words: multisensory integration, exogenous spatial cuing, temporal binding window, crossmodal response enhancement

Multisensory integration and exogenous spatial attention:
a time-window-of-integration analysis

INTRODUCTION

The response of a multisensory neuron to stimulation in two (or more) sensory modalities can exceed that observed of either stimulus in isolation. This has been found at a variety of brain sites including subcortical convergence-zones such as the superior colliculus, and also within cortical areas such as ectosylvian cortex in cats, or superior, temporal, parietal, and premotor cortex in monkeys. Specifically, the absolute number of impulses (spikes) of a neuron, registered within a fixed time interval after bimodal stimulus presentation, is larger than in the unimodal conditions provided that the two stimuli occur fairly close in time and space (Stein & Meredith, 1993). This classic result of multisensory research, often expressed in terms of so-called spatial and temporal "rules of integration", has given rise to the concept of a "spatio-temporal time window of integration" (Meredith, 2002). In the behavioral domain, a corresponding concept is typically referred to as "temporal binding window": even with stimulus onset asynchronies (SOAs) of several hundred milliseconds, stimuli from different sensory modalities are "bound" to signal the occurrence of a joint event that leads crossmodal information to trigger faster responses or to lower detection and discrimination thresholds (see Wallace & Stevenson, 2014, for a recent review).

In contrast to studies based on data from single neurons with their overlapping unimodal receptive fields, the possible sources of such behavioral and perceptual effects have been discussed controversially. Early on it has been well documented that occurrence of an irrelevant and non-predictive sound facilitates motor responses to a subsequent target light appearing nearby (e.g., McDonald, Teder-Sälejärvi, & Hillyard, 2000). This "exogenous spatial cuing effect" (Driver & Spence, 2004) is interpreted as involuntary, crossmodal re-orienting (shift) of visual attention. There is overwhelming evidence now from numerous studies, both behavioral and electrophysiological, that a sound attracts visual attention to its location even when it provides no information about where a subsequent visual target may occur, or even when no visual stimuli are presented at all (for the latter, see e.g., Brang et al., 2015). This shift of ‘visual attention’ manifests itself, amongst others, in terms of speeding detection, improving discrimination of a visual target, or enhancing perceived contrast (see Hillyard, Störmer, Feng, Martinez, & McDonald, 2016, for a recent review). The central issue is whether the distinction between multisensory integration and exogenous spatial cuing is just terminological or is based on distinct neural underpinnings (Macaluso, Frith, & Driver, 2000, 2001). In their review chapter, Spence and colleagues (Spence, McDonald, & Driver, 2004) speculated that exogenous spatial cuing might be produced by “backprojection influences from putative supramodal ‘attention’ structures (say, in parietal and/or superior temporal cortex) to representations specific to the target modality (e.g. occipital cortex), with these backprojections being induced by an event in the cue modality even when presented on its own”, whereas multisensory integration effects “might be characterized as arising instead only when feedforward projections from each stimulated modality meet at the multimodal level” (ibid, p. 311). However, neurophysiological studies have shown that two association cortical areas (anterior ectosylvian sulcus and the rostral lateral suprasylvian sulcus) have crucial roles to play in multisensory integration occurring in superior colliculus (Alvarado, Stanford, Vaughan, & Stein, 2007), blurring the proposed distinction to some degree.

The temporal profiles of multisensory integration and exogenous spatial cuing effects clearly differ: as observed in Van der Stoep, Spence, Nijboer, and Van der Stigchel (2015, p. 21), the facilitative effect of exogenous shifts of attention is typically most pronounced when SOA between presentation of the auditory and visual stimulus is about 100 to 300 ms, whereas behavioral benefits of multisensory integration are often most pronounced when auditory and visual stimuli are presented in close temporal alignment (SOAs between 0 and ±50 ms) or in “physiological synchronicity”, that is, taking the peripheral transition times of different modalities into account (Diederich & Colonius, 2004a). Naturally, these numbers should be taken with a grain of salt because the actual values will always depend on the specifics of the experimental (stimulus and task) conditions. Nevertheless, the main point to be exploited in the modeling to follow below is that exogenous shifts of attention take some time to become effective, whereas multisensory integration has no lower limit of SOA to occur (but, obviously an upper limit). Van der Stoep and colleagues (Van der Stoep et al., 2015) set out to separate multisensory integration from crossmodal exogenous spatial attention by presenting participants exactly the same auditory and visual stimuli in two different tasks. In the Integration block, they were instructed to respond to the onset of the first target stimulus that they detected (A or V). The instruction for the Cuing block was to respond only to the onset of the visual targets. The targets could appear at one of three locations: left, center, and right. The participants were instructed to respond only to lateral stimuli, targets appearing at the central location serving as 'catch' trials. The authors then compared the two types of tasks with respect to the crossmodal response enhancement (MRE), measured as difference between median unimodal and crossmodal reaction times. Without going into further details of their procedures (but see Van der Stoep et al., 2015), they found MRE in the Integration block only when the auditory and visual stimuli were presented in close spatial and temporal alignment, i.e., at the 0 and 50 ms SOA (with the auditory stimulus always presented first). Moreover, they concluded from further analysis that (i) at SOAs of 100 and 200 ms, MRE is likely driven by the effects of crossmodal exogenous spatial attention, alerting, and response preparation but cannot be explained by multisensory integration, and (ii) that at SOA of 50 ms both types of processing may have occurred.

While these results are very interesting, some aspects of Van der Stoep et al. (2015) require further scrutiny. First, in order to elicit both multisensory integration and spatial cuing participants had to undergo two different experimental paradigms: a redundant signals task ("respond to stimuli from both modalities") in the Integration block and a focused attention task ("only respond to stimuli from the target modality") in the Cuing block. This leaves open the question of whether both integration and cuing may, in principle, occur within one and the same trial. Second, whereas measuring the strength of multisensory integration via the amount of violation of the race model inequality (RMI) is common practice for the redundant signals task, its use in the focused attention task is rather dubious for the following reason: RMI provides a lower bound for MRE assuming that, in a given trial, the response is elicited by whichever stimulus is detected first. However, in the focused attention task, if the non-target is the winner of the race, the response must be inhibited, i.e. postponed up until the target is detected.

TIME WINDOW OF INTEGRATION (TWIN) MODEL

Results from studies of the temporal profile of crossmodal MRE can conveniently be described within the framework of the time window of integration (TWIN) model developed over the last 15 years by the authors (Colonius & Diederich, 2004; Diederich & Colonius, 2004b, 2015). After introducing the basic assumptions, an extension of the model incorporating both integration and spatial cuing effects is presented.

Basic TWIN model framework

In order to predict the spatiotemporal effects on saccadic or manual RT observable in a simple detection task, the model postulates that a crossmodal (audiovisual) stimulus triggers a race mechanism in the very early, peripheral sensory pathways (*first stage*), followed by a compound stage of converging sub-processes that comprise neural integration of the input and preparation of a response. This *second stage* is defined by default: it includes all subsequent, possibly temporally overlapping processes that are not part of the first stage. A central assumption concerns the temporal configuration needed for crossmodal interaction to occur:

> TWIN assumption: crossmodal interaction occurs only if the peripheral processes of the first stage all terminate within a given temporal interval, the

time window of integration.

Thus, the window can be considered a filter determining whether afferent information delivered from separate sensory organs is registered close enough in time to trigger multisensory integration. Given the nature of the task –simple stimuli and response via button press (no choice) or an eye movement toward the target– suggests that this first, peripheral stage is not influenced by the spatial distance between, e.g., the visual and auditory stimulus. Importantly, this implies that, in the TWIN model, possible crossmodal effects of the temporal configuration can be separated from spatial effects of the experimental setup. The amount of crossmodal interaction manifests itself in an increase, or decrease, of second stage processing time. While this amount no longer depends on SOA of the stimuli, temporal tuning of the interaction nevertheless is made possible because the *probability of integration* is modulated by the SOA value. The formal version of the model makes these assumptions explicit:

The race in the first stage is represented by two statistically independent, non-negative random variables V and A, standing for the peripheral visual and auditory processing times, respectively. With τ as SOA value and ω as (integration) window width parameter, the TWIN assumption for multisensory integration to occur, denoted by I, is

$$I = \{A + \tau < V < A + \tau + \omega\},$$

for the case of visual stimulus as target and auditory as non-target. The starting point of the visual stimulus is arbitrarily defined as physical zero time point; thus, negative τ values mean that the auditory non-target is presented before the visual target. The probability of integration to occur, $P(I)$, is a function of both τ and ω, and it can be calculated once distribution functions for A and V have been specified. Note that, in particular, multisensory integration will not occur if the target stimulus wins in the first stage.

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

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

Here, the assumption is that the first stage terminates with processing of the target, so $S_1 = V$. The complementary event to I is denoted by I^c, μ is short for $\mathrm{E}[S_2|I^c]$, $P(I)$ the probability of integration, and Δ_I stands for $\mathrm{E}[S_2|I^c] - \mathrm{E}[S_2|I]$. The term $P(I) \times \Delta_I$ is a measure of the expected amount of crossmodal interaction in the second stage, with positive Δ_I values corresponding to facilitation, negative ones to inhibition. Obviously, event I cannot occur in the unimodal visual condition, thus expected reaction time is

$$\mathrm{E}[RT_V] = \mathrm{E}[V] + \mathrm{E}[S_2|I^c] = \mathrm{E}[V] + \mu.$$

Defining MRE as difference between uni- and crossmodal condition,

$$\begin{aligned}\mathrm{MRE} &= \mathrm{E}[RT_V] - \mathrm{E}[RT_{V\tau A}] \\ &= P(I) \times \Delta_I,\end{aligned}$$

with $P(I)$ depending on both SOA and window width ω and Δ_I only depending on crossmodal aspects of the experiment like, in particular, the spatial distance between the stimuli. In empirical studies probing various aspects of TWIN (Colonius, Diederich, & Steenken, 2009; Diederich & Colonius, 2008a, 2008b, 2007a, 2007b), the peripheral processing times have been assumed to follow exponential distributions. Moreover, adding a Gaussian component as second stage processing time results in predicting RT distributions in the form of (a mixture of) ex-Gaussian distributions that are quite common in RT modeling (Luce, 1986).

Extended TWIN model framework

In order to accommodate a (spatial) cuing mechanism in the TWIN model, the following assumption is added:

> Cuing assumption: a (spatial) cuing mechanism –resulting in a speedup of second stage processing time– occurs whenever peripheral processing of the non-target stimulus terminates "early enough" before peripheral target stimulus processing is finished.

The intuition here is that a cue needs a sufficient head start to "shift visual attention" to its location and to enable enhanced processing. In analogy to multisensory integration, the occurrence of cuing and the amount of its effect are separated such that the cuing effect becomes manifest in a speedup of second stage processing. Note that only the case of a facilitating cue is considered here, but a more general formulation could easily be introduced. The specification of "early enough" is by introducing an additional model parameter, denoted by α, so that the event W that cuing occurs is specified as follows:

$$W = \{A + \tau + \alpha < V\},$$

with $\alpha > 0$. Note that large α-values make the probability of cuing to occur small and vice versa. Prediction for mean crossmodal RT then changes into

$$\begin{aligned}\mathrm{E}[RT_{V\tau A}] =& \mathrm{E}[S_1] + P(I \cap W)\,\mathrm{E}[S_2 \mid I \cap W] \, + P(I^c \cap W)\,\mathrm{E}[S_2 \mid I^c \cap W] \\ &+ P(I \cap W^c)\,\mathrm{E}[S_2 \mid I \cap W^c] + P(I^c \cap W^c)\mathrm{E}[S_2 \mid I^c \cap W^c],\end{aligned} \tag{2}$$

with $P(I \cap W)$ denoting the probability that, in a given trial, both integration and cuing occur, $P(I^c \cap W)$ the probability that integration does not occur and cuing does occur, and so on. Using the fact that these probabilities sum up to 1, crossmodal mean can be expressed as

$$\mathrm{E}[RT_{V\tau A}] = \mathrm{E}[RT_V] + \mu - [P(I \cap W)\Delta_{IW} + P(I^c \cap W)\Delta_W + P(I \cap W^c)\Delta_I], \tag{3}$$

with $\mu = \mathrm{E}[S_2 \mid I^c \cap W^c]$ denoting second stage processing time when neither integration nor cuing takes place, $\Delta_{IW} = \mu - \mathrm{E}[S_2 \mid I \cap W]$, $\Delta_W = \mu - \mathrm{E}[S_2 \mid I^c \cap W]$, and $\Delta_I = \mu - \mathrm{E}[S_2 \mid I \cap W^c]$. Thus, MRE becomes a weighted sum of three different facilitation effects

$$\begin{aligned}\text{MRE} &= \text{E}[RT_V] - \text{E}[RT_{V\tau A}] \\ &= P(I \cap W)\Delta_{IW} + P(I^c \cap W)\Delta_W + P(I \cap W^c)\Delta_I,\end{aligned}$$

where Δ_{IW} represents the combined effect of integration and cuing, Δ_W the effect of cuing without integration, and Δ_I the effect of integration without cuing. In order to apply this modeling framework to a crossmodal RT experiment, the Δ parameters have to be adapted to the experimental conditions, e.g., whether crossmodal stimuli occur at coincident or disparate locations. Moreover, specific hypotheses can be formulated, e.g., whether or not integration and cuing have additive effects such that, e.g., $\Delta_{IW} = \Delta_I + \Delta_W$.

TWIN Model specification

The first step in deriving predictions from the TWIN model is to compute the probability of integration and cuing as a function of the SOA (τ) values used in the experiment. Second, Δ values have to be assigned to the various spatial conditions under which RTs have been collected.

Probability of integration and cuing. We first consider the probability of both integration and cuing occurring, $P(I \cap W)$, in a given trial. Peripheral processing times A and V are assumed to be exponentially distributed with parameters λ_A and λ_V, respectively. From $I = \{A + \tau < V < A + \tau + \omega\}$ and $W = \{A + \tau + \alpha < V\}$, we conclude

$$P(I \cap W) = P(A + \tau + \alpha < V < A + \tau + \omega). \tag{4}$$

In this case, $\alpha < \omega$ follows because both α and ω are nonnegative. Given that all SOA (τ) are negative or zero, three different cases must be considered in computing the above probability: (i) $\tau + \alpha \geq 0$; (ii) $\tau + \alpha < 0$ and $\tau + \omega \geq 0$, and (iii) $\tau + \alpha < 0$ and $\tau + \omega < 0$. For case (i), using the assumption that A and V are stochastically independent,

$$\begin{aligned}P(I \cap W) &= \int_0^\infty f_A(a)[F_V(a + \tau + \omega) - F_V(a + \tau + \alpha)]\,\mathrm{d}a \\ &= \frac{\lambda_A}{\lambda_A + \lambda_V}(-1)\exp[-\lambda_V(\alpha + \tau) - \lambda_V(\omega + \tau)]\{\exp[\lambda_V(\alpha + \tau)] - \exp[\lambda_V(\omega + \tau)]\}.\end{aligned}$$

Cases (ii) and (ii) are obtained in the same way resulting in

$$P(I \cap W) = \frac{\lambda_A}{\lambda_A + \lambda_V}(1 - \exp[-\lambda_V(\omega + \tau)]) + \frac{\lambda_V}{\lambda_A + \lambda_V}(1 - \exp[\lambda_A(\alpha + \tau)]),$$

for case (ii), and

$$P(I \cap W) = \frac{\lambda_V}{\lambda_A + \lambda_V}(\exp[\lambda_A(\omega + \tau)] - \exp[\lambda_A(\alpha + \tau)])$$

for case (iii).

The remaining probabilities, $P(I^c \cap W)$, and $P(I \cap W^c)$, can be computed similarly, and $P(I^c \cap W^c)$ follows because the probabilities have to sum up to 1 (details are available on request from the authors' MatLab© code).

Different versions of the second stage of TWIN. The different spatial conditions of the experiment fully described below are: visual and auditory stimulus are presented (i) either coincident (same position in space) or disparate (different positions in space) and (ii) either in the horizontal or the vertical axis. A fully saturated model would allow distinct Δ-parameters in the response enhancement term for each of these four possible configurations,

$$\text{MRE} = P(I \cap W)\Delta_{IW} + P(I^c \cap W)\Delta_W + P(I \cap W^c)\Delta_I,$$

resulting in $3 \times 4 = 12$ Δ-parameters for second stage processing. Moreover, the probability of integration and/or cuing, depending on window width (ω) and head start parameter (α), could in principle also be modulated by spatial configuration resulting in a redundancy of parameters that would be difficult to identify.

Introducing different parameter restrictions renders a number of model versions amenable to testing, three of which are considered here. In the first model, no cuing mechanism is postulated, so RT differences between spatial stimulus configurations are assumed to be due to multisensory integration alone. In the second, both spatial cuing and integration may occur, while the third version allows different time window widths for vertical vs. horizontal presentation, multisensory integration, but no spatial cuing. A more detailed description follows.

Model 1: Integration without cuing. Here each spatial configuration may have a different effect on integration, but cuing has no effect at all. Thus, response enhancement becomes

$$\text{MRE}_{ij} = P(I \cap W^c)\Delta_{ij},$$

with $i = h, v$ for horizontal and vertical configuration and $j = c, d$ for coincident and disparate configuration, respectively. Other parameters are λ_A, λ_V for auditory and, respectively, visual peripheral processing time, μ for second stage processing without enhancement, and ω for window width, resulting in 8 parameters in total.

Model 2: Integration and horizontal/vertical-specific cuing. In this model, integration occurs specific to coincident vs. disparate presentation (Δ_c and Δ_d), spatial cuing occurs (Δ_W) with vertical/horizontal-specific head start (α_v and α_h) :

$$\text{MRE}_j = P(I \cap W)\Delta_{IW} + P(I^c \cap W)\Delta_W + P(I \cap W^c)\Delta_j, \tag{5}$$

with $j = c, d$. This implies 9 parameters in total.

Model 3: Integration but no cuing, and horizontal/vertical-specific time windows. In this model, integration occurs specific to coincident vs. disparate presentation (Δ_c and Δ_d), no spatial cuing is assumed but time window is vertical/horizontal-specific (ω_v and ω_h)

$$\text{MRE}_j = P(I \cap W^c)\Delta_j,$$

with $j = c, d$. Total number of parameters is 7 seven.

Given the number of parameters and the nonlinear structure of the TWIN model, discriminating between these different model versions may be difficult. The following experiment is meant as a first step to illustrate the general approach.

EXPERIMENT

We want to investigate whether the temporal profile of response enhancement in a crossmodal task is consistent with the separation of facilitation effects into a part due to multisensory integration and another part due to a shift of attention/spatial cuing. Since elicitation of the latter requires presentation of a cue prior to the stimulus, a focused attention paradigm with a visual target and an auditory non-target with varying SOA values was chosen. Moreover, in order to elicit possibly different cuing effects we set up visual-auditory spatial configurations with either horizontal or vertical disparity. A number of studies have shown that, depending on SOA, response facilitation with vertically arranged stimuli may differ from that to horizontal ones, presumably due to differing auditory localization mechanisms (Frens, van Opstal, & van der Willigen, 1995)

Participants. Four students, aged 19 to 21, 1 female, from Jacobs University Bremen, were recruited as voluntary participants. They were paid or received credit points required for their studies. All had normal or corrected-to-normal vision, normal hearing, and were right-handed (self-description, Coren's Lateral Preference Inventory, 1993). Participant gave their written informed consent prior to their inclusion in the study and the experiment has been conducted according to the principles expressed in the Declaration of Helsinki. They were screened for their ability to follow the experimental instructions. Approval for this study was granted by the Academic Integrity Committee of Jacobs University Bremen.

Stimuli and apparatus. Fixation point was an LED (25 mA, 5.95 mcd) located in the medial line at a distance of 90 cm. Auditory stimuli were white noise bursts (59 dBA), generated by two speakers (Canton Plus XS). Visual stimuli were red LEDs (25 mA, 3.3 mcd) located on top of the speakers. There were two different spatial arrangements: for the horizontal arrangement, the speakers were placed at 20° to the left and right of the fixation LED at the height of participants' ear level at a distance of 120 cm; for the vertical arrangement, the speakers were placed at the medial line 15° above and below the fixation point. Note that the distance between visual and auditory stimuli presented at the same position ("coincident") was always 0°; for stimuli presented at opposite positions ("disparate"), the distance between two stimuli amounted to 40° for the horizontal and to

30° for the vertical arrangement. Figure 1 shows the stimulus configuration schematically. Response devices were two reflective optical sensors mounted on the table in front of the participant.

Experimental task and procedure. Participants were seated in a completely darkened, sound attenuated room with the head positioned on a chin rest, elbows and lower arms resting comfortably on a table. The participant began every experimental session with a 10 min of dark adaptation during which the measurement system was adjusted and calibrated. Each trial started with the appearance of the fixation point of random duration (800–1500 ms). With the off-set of the fixation LED the visual target stimulus was turned on. The onset of auditory non-targets was shifted by a stimulus onset asynchrony (SOA) from -250 to 0 ms in steps of 50 ms. The non-targets were turned off simultaneously with the visual target stimulus which lasted for 500 ms. Stimulus presentation was followed by a break of 2 s in complete darkness before the next trial began, indicated by the onset of the fixation LED. Participants were instructed to put their left and right index fingers on the sensor buttons related to the spatial position of the visual target and to lift the respective finger as quickly and accurately as possible ignoring any auditory non-targets (focused attention paradigm). Such spatial stimulus-response arrangement is considered necessary for spatial cuing effects to occur (Lee & Charles, 2017) . Specifically, for the horizontal arrangement the right (left) hand button was associated with the target appearing on the right (left) side of the participant. For the vertical arrangement the right (left) hand button was associated with the target appearing above (below) the fixation LED. The visual target appeared in combination with the auditory non-target in either coincident or disparate position. One experimental block consisted of 176 trials, 168 bimodal and 8 unimodal (visual target alone) with either a horizontal or a vertical orientation. Each participant performed a total of 28 experimental blocks, four of them presented in one session which lasted for about an hour. Each session consisted of two horizontal and two vertical blocks in alternating order. Each participant was engaged for seven hours over the course of two weeks and completed a total of $4,928$ experimental trials.

RESULTS

Data screening and preprocessing

RTs were screened for anticipation errors (RT < 80 ms), misses (RT > 500 ms), left/right directional errors, and non-responses (by not lifting the finger properly). Overall error rates (in %) for horizontal (vertical) orientation were: 10 (12) for Participant S1; 4 (9) for S2; 2(4) for S3; and 20(17) for S4.

Mean RTs

While participants differed slightly in their overall response speed, their mean RTs across SOA values from -250 to 0 ms exhibit very similar behavior for the four different combinations of spatial conditions, i.e., horizontal-vertical and coincident-disparate. The following general pattern (with very few exceptions) emerges from inspecting Figure 2: (i) facilitation is always occurring (except for 2 out of 96 conditions), and it decreases the later the auditory stimulus is given relative to the visual target; (ii) strength of facilitation is the largest under horizontal-coincident presentation, followed by horizontal-disparate, vertical-coincident, and finally, vertical-disparate presentations. Thus, both spatial features, the horizontal-vertical and coincident-disparate distinction, have an effect on facilitation, with the former clearly dominating the latter. Moreover, the effect of coincident vs. disparate presentation remains more or less invariant across the different SOA values. There is also an impression of further speed-up beyond the 250 ms auditory lead, as actually observed in other studies from our lab (e.g., Diederich & Colonius, 2008a).

Parameter estimation for TWIN

Performance of Models 1–3 was probed separately for each participant. Seven to nine parameters, depending on the model, were estimated from the means of 24 bimodal conditions: 2 spatial arrangements (horizontal/vertical) × 2 incidence levels (coincident/disparate) × 6 SOA values. Common to all models are parameters λ_A and λ_V for peripheral processing time of auditory and visual stimuli, respectively, and μ presenting mean time of second stage processing when cuing or integration effects are absent. The remaining model specific parameters have been described above and are listed in Table1.

All parameters were estimated by minimizing the following χ^2 statistic,

$$\chi^2 = \sum_{n=1}^{6} \sum_{k=1}^{2} \sum_{j=1}^{2} \left(\frac{RT_{obs}(j,k,n) - RT_{pred}(j,k,n)}{\sigma_{RT_{obs}(j,k,n)}} \right)^2 \tag{6}$$

using optimization routine BADS (Acerbi & Ji, 2017) implemented in Matlab R2017a. Here, $RT_{obs}(j,k,n)$ and $RT_{pred}(j,k,n)$ are, respectively, the observed and predicted values of mean RT to visual-auditory stimuli, presented in positions (coincident, $j = 1$; disparate, $j = 2$) and spatial arrangement (horizontal, $k = 1$; vertical, $k = 2$) with SOA (referred to by n, $n = 1, \ldots, 6$); $\sigma_{RT_{obs}(j,k,n)}$ are the respective standard errors (standard deviation divided by the square root of the number of observations). For the estimation routine, the following restrictions were made: λ_V and λ_A between 5 and 300 ms; μ between 50 to 500 ms; ω between 50 and 250 ms; moreover, α and Δ values were restricted between 0 and 200 ms.

TWIN model account

Parameter estimates for all participants and models (M1–M3) are listed in Table 1. Some parameter patterns are invariant across all participants and models: (i) peripheral processing for auditory stimuli takes longer than for visual stimuli ($1/\lambda_A > 1/\lambda_V$), (ii) the effect of coincident stimulation is stronger than disparate stimulation (compare Δ values); (iii) second stage processing (μ) is between 250 and 400 ms. Moreover, there is high consistency across all participants with respect to parameters in the three model versions: (i) in models without cuing, time window width is equal or close to the upper bound of 250 ms (M1), and window widths for horizontal presentations are longer than for vertical ones (M3); (ii) cuing parameter α is about 100 ms for horizontal and zero for vertical presentations (in 3 out of 4 participants). The fact that peripheral auditory processing was estimated to take longer than visual across all participants seems not consistent with the fact that transmission delays are typically estimated faster for auditory than visual. Note, however, that first-stage peripheral processing time in TWIN comprises more than (neurophysiological) transmission delays, referring to the entire afferent chain up until a first interaction between visual and auditory processing is possible, i.e., purportedly at early cortical processes.

Goodness of fit information is listed in the two final rows of Table 1. For comparing models with different numbers of parameters, we used the Akaike information criterion (AIC) here defined as AIC $= \chi^2 + k \cdot (k+1) - 2 \cdot df$, where k is number of parameters and degrees of freedom $df = N - k$ with N the number of observations (Kenny, 2015). Model fits are consistent across all 4 participants. Specifically, the model allowing for both integration and cuing (Model 2) is the clear winner both in terms of the χ^2 statistic (Equation 6) and when taking the number of parameters into account (AIC value). When the cuing mechanism is replaced by distinguishing vertical/horizontal from coincident/disparate spatial effects (four Δ-values) im Model 1, the fit degrades somewhat. Finally, the fit suffers even more when cuing is replaced by allowing different time windows for horizontal and vertical presentations while keeping the coincident/disparate effect in Δ (Model 3). Figure 2 shows predictions of the best-fitting model (M2) for all participants. As the 95% -confidence bands for mean RTs indicate, the model gives a good account of the data. It captures the overall pattern for the different spatial conditions (horizontal/vertical and coincident/disparate) and also accounts for (minor) individual differences.

Given estimates of all model parameters, the extended TWIN model not only yields estimates of the facilitation effects under each of the experimental conditions but also allows to see how the effect of cuing and/or integration evolves across different SOA values. As depicted in Figure 3, for all participants and with SOA going from -250 to 0 ms, (i) the probability of cuing decreases under both horizontal and vertical configurations (dash-dotted/blue curves); (ii) the probability of both cuing and integration to occur tends to first increase and then decrease (dashed/red curves); (iii) the probability of integration occurring without cuing follows a similar shape (dotted/black curves) with the following exception: it remains at probability zero for those conditions where $\alpha_h = 0$ (see Table 1). This is easy to understand: in those cases we have $W^c = \{A + \tau + 0 > V\}$ (no cuing), which implies that the probability of “no cuing but integration”, $P(I \cap W^c)$, must be zero because integration requires $\{A + \tau < V\}$. Finally, the solid (cyan) curves, monotonically increasing towards about 0.7, indicate the probability that neither cuing nor integration occur. Note that all these curves, while not being directly observable, give an excellent insight into how the two purported mechanisms are at work at different levels of SOA.

Finally, an important feature of the TWIN model is that it teases apart the probabilities of cuing/integration from the numerically quantified effects, captured by Δ-values as expressed, e.g., in Equation 5 for Model M2. While neither the probabilities nor the Δ-values are observable, their combined effect results in the observed amount of facilitation (MRE). Specifying which aspects of the experimental setup influence the probabilities, on the one hand, and the Δ-values, on the other hand, makes predictions of TWIN empirically testable.

DISCUSSION AND CONCLUSION

Facilitating effects of a non-target auditory stimulus on the speeded response to a subsequent visual target stimulus are well-documented (see, e.g., Driver & Spence, 2004, for a review), but the source of this “exogenous spatial cuing” effect is somewhat controversial. Specifically, it is not clear whether facilitation is due to ‘crossmodal links in exogenous spatial attention’ or should be considered as outcome of a multisensory integration mechanism, as typically observed at the single cell level (e.g., McDonald, Teder-Sälejärvi, & Ward, 2001). Here, we extend the time-window-of-integration (TWIN) model (Colonius & Diederich, 2004; Diederich & Colonius, 2004b) to combine both sources of facilitation within the same formal framework. A (spatial) cuing mechanism is added that operates whenever peripheral processing of a non-target stimulus terminates “early enough” before target stimulus processing. Multisensory response enhancement, i.e., RT facilitation relative to unimodal response time, occurs in a second-stage whenever multisensory integration and/or cuing are induced by a probabilistic time window mechanism in the first-stage.

The model was illustrated by fitting it to data from a focused attention task with a visual target and an auditory non-target presented at horizontally or vertically varying positions with SOA values ranging from -250 to 0 ms (auditory lead). We find facilitation across the entire range of SOAs, but the size of the effect depends on the specific spatial configuration of target and non-target: it is stronger when both stimulus positions are varied horizontally rather than vertically. With both stimuli presented at the same position (coincident), facilitation was stronger than for stimuli presented at different positions (disparate), and this effect was less salient than for the horizontal/vertical distinction. These findings are consistent with effects observed in saccadic eye movements with respect to both spatial coincidence and the horizontal/vertical dimension (Frens et al., 1995) and also replicate previous data from our lab where, however, the impact of prior knowledge (Diederich, Colonius, & Kandil, 2016) and auditory maskers (Colonius et al., 2009) had been in the focus of investigation, respectively. Here, different model versions were fit to the data depending on whether or not multisensory integration was postulated as the sole mechanism underlying RT facilitation. The model version allowing both integration and spatial cuing to occur within a trial (Model 2) delivered the best account for the 96 mean RT data collected from four participants.

Comparing our data from the horizontal spatial condition with those in the Cuing block of Van der Stoep et al. (2015), the only condition that parallels our focused attention paradigm, reveals a high degree of consistency for both coincident and disparate presentations: for SOA $= -50, -100, -200$ ms, when the non-target auditory stimulus gets a lead, responses accelerate monotonically with a total speed-up of about 80 ms for coincident presentations and somewhat less for disparate ones. But on closer scrutiny, there are also some differences. Whereas Van der Stoep and colleagues find little or no facilitation at SOA = 0, we do in 3 out of 4 participants. On the other hand, computation of the exact amount of facilitation critically depends on the unimodal reference point, usually the minimum of the mean visual- and/or auditory-only condition, as a function of the specifics of the experimental conditions like blocked or mixed presentations, catch trials etc., that were not completely identical in both studies. A potentially more important discrepancy concerns the spatial alignment effect: while Van der Stoep et al. (2015) report a significant increase in the difference between coincident and disparate conditions across SOAs, in our data the displacement of RTs from coincident to disparate configurations remains more or less invariant for both horizontal and vertical conditions (see Figure 2). Comparing the Cuing block data with those in the Integration block, the authors hypothesize that, in the Cuing block, there was no integration in the 100 and the 200 ms SOA conditions but, rather, that facilitation was the result of combined effects of crossmodal exogenous spatial attention, alerting, and temporal preparation effects, while at SOA = 0 the observed MRE was, at least in part, a result of multisensory integration. In contrast, in the TWIN model the probability that both integration and cuing occur at SOA = 0, 100, or 200 is nearly always larger than the probability that only cuing occurs, in particular for the horizontal configuration.

To summarize the main findings of this paper: first, while there were some variations between Van der Stoep et al. (2015) and our data, we also find high consistency for most SOA conditions but, second, the interpretation of the cause for facilitation, cuing and/or integration, deviates significantly. Third, the specific contribution of the extended TWIN model is to combine both multisensory integration and spatial cuing in a single process. Notably, the model allows to quantify the effect of either mechanism in facilitating the response within one and the same trial. Once model parameters have been estimated, the model can predict MRE under various, previously untested spatiotemporal configurations of the stimuli. Specifically, it is possible, at least in principle, to probe whether the conclusions drawn in Van der Stoep et al. (2015) based on comparing data from the Integration and the Cuing block would actually hold up when the fitting the TWIN model to their data.

Although the model fits presented here show a consistent advantage for the assumption that both integration and cuing co-exist in bringing about crossmodal facilitation of reaction time and that both mechanisms may occur in one and the same trial, a larger data set and more advanced model selection techniques may be required to give a more definite answer to which model version is closest to the truth. As suggested by a reviewer, an informative modification might be an experiment using both within-modal and crossmodal stimuli: the former should only produce cuing in contrast to the latter. However, as reported by Schröter and colleagues investigating the redundant signals task using within-modal stimuli without SOAs (Schröter, Frei, Ulrich, & Miller, 2009; Schröter, Fiedler, Miller, & Ulrich, 2011), the combination of two or more unimodal visual or auditory stimuli, may produce unexpected results. Specifically, facilitation of RT depended on the number of percepts, rather than the number of stimuli. Therefore, an experiment to compare results from crossmodal to within-modal stimuli within a focused attention task will require a judicious choice of appropriate stimulus parameters.

Acknowledgments

AD was supported by grant DI 506/12-1 (German Science Foundation/DFG). HC was supported by a grant from German Science Foundation/DFG (SFB/TRR31) and Cluster of Excellence (DFG) “Hearing4all” of Oldenburg University.

References

Acerbi, L., & Ji, W. (2017). Practical Bayesian optimization for model fitting with Bayesian Adaptive Direct Search. In I. Guyon et al. (Eds.), Advances in Neural Information Processing Systems 30 (pp. 1836–1846). Curran Associates, Inc.

Alvarado, J., Stanford, T., Vaughan, J., & Stein, B. E. (2007). Cortex mediats multisensory but not unisensory integration in superior colliculus. The Journal of Neuroscience, 27(47), 12775–12786.

Brang, D., Towle, V., Suzuki, S., Hillyard, S., Di Tusa, S., Dai, Z., . . . Grabowecky, M. (2015). Peripheral sounds rapidly activate visual cortex: evidence from electrocorticography. Journal of Neurophysiology, 114, 3023–3028.

Colonius, H., & Diederich, A. (2004). Multisensory interaction in saccadic reaction time: A time-window-of-integration model. J Cogn Neurosci, 16, 1000–1009.

Colonius, H., Diederich, A., & 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, 3–4(177–184).

Diederich, A., & Colonius, H. (2004a). Bimodal and trimodal multisensory enhancement: effects of stimulus onset and intensity on reaction time. Perception and Psychophysics, 66, 1388–1404.

Diederich, A., & Colonius, H. (2004b). Handbook of multisensory processes. In G. Calvert, C. Spence, & E. Stein B. (Eds.), (p. 395-408). Cambridge: MIT Press.

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

Diederich, A., & Colonius, H. (2007b). Why two "distractors" are better then one: Modeling the effect on non-target auditory and tactile stimuli on visual saccadic reaction time. Experimental Brain Research, 179, 43–54.

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

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

Diederich, A., & 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., Colonius, H., & Kandil, F. I. (2016). Prior knowledge of spatiotemporal configuration facilitates crossmodal saccadic response : A twin analysis. Experimental Brain Research, 234(7), 2059–2076.

Driver, J., & Spence, C. (2004). Crossmodal spatial attention: evidence from human performance. In C. Spence & J. Driver (Eds.), Crossmodal space and crossmodal attention (pp. 179–220). New York, NY: Oxford University Press.

Frens, M., van Opstal, A., & van der Willigen, R. (1995). Spatial and temporal factors determine auditory-visual interactions in human saccadic eye movements. Perception & Psychophysics, 57(6), 802–816.

Hillyard, S., Störmer, V., Feng, W., Martinez, A., & McDonald, J. (2016). Cross-modal orienting of visual attention. Neuropsychologia, 83, 170–178.

Kenny, D. (2015, November). Measuring model fit. Retrieved June 6, 2018, from `http://davidakenny.net/cm/fit.htm`

Lee, J., & Charles, S. (2017). On the spatial specificity of audiovisual crossmodal exogenous cuing effects. Acta Psychologica, 177, 78–88.

Luce, R. (1986). Response times: Their role in inferring elementary mental organization. New York, NY: Oxford University Press.

Macaluso, E., Frith, C. D., & Driver, J. (2000). Modulation of human visual cortex by crossmodal spatial attention. Science, 289, 1206–1208.

Macaluso, E., Frith, C. D., & Driver, J. (2001). A reply to McDonald, J.J., Teder-Sälejärvi, W.A., and Ward, L.M. Multisensory integration and crossmodal attention effects in the human brain. Science, 292, 1791.

McDonald, J. J., Teder-Sälejärvi, W. A., & Hillyard, S. A. (2000). Involuntary orienting to sound improves visual perception. Nature, 407, 906–908.

McDonald, J. J., Teder-Sälejärvi, W. A., & Ward, L. M. (2001). Multisensory integration and cross-modal attention effects in the human brain. Science, 292, 1791.

Meredith, M. (2002). On the neuonal basis for multiensory convergence: a brief overview. Cognitive Brain Research, 14, 31–40.

Schröter, H., Fiedler, A., Miller, J., & Ulrich, R. (2011). Fusion prevents the redundant signals effect: evidence from stereoscopically presented stimuli. Journal of Experimental Psychology: Human Perception and Performance, 37(5), 1361–1368.

Schröter, H., Frei, L., Ulrich, R., & Miller, J. (2009). The auditory redundant signals effect: An influence of number of stimuli or number of percepts? Attention, Perception, & Psychophysics, 71(6), 1375–1384.

Spence, C., McDonald, J., & Driver, J. (2004). Exogenous spatial-cuing studies of human crossmodal atttention and multisensory integration. In C. Spence & J. Driver (Eds.), Crossmodal space and crossmodal attention (pp. 277–320). New York, NY: Oxford University Press.

Stein, B. E., & Meredith, M. (1993). The merging of the senses. Cambridge, MA: MIT Press.

Van der Stoep, N., Spence, C., Nijboer, T., & Van der Stigchel, S. (2015). On the relative contributions of multisensory integration and crossmodal exogenous spatial attention to multisensory response enhancement. Acta Psychologica, 162, 20–28.

Wallace, M., & Stevenson, R. (2014). The construct of the multisensory temporal binding window and its dysregulation in developmental disabilities. Neuropsychologia, 64, 105–123.

Figure captions

Figure 1 Stimulus configuration and time line. A: vertical spatial arrangement with disparate stimuli; B: vertical with coincident stimuli; C: horizontal spatial arrangement with disparate stimuli; D: horizontal with coincident stimuli. Note that the target was presented at the top or bottom (not shown in the figure) and left or right (not shown). FIX indicated the fixation point presented at the beginning of each trial. E: time line for one trial. Note that the graph does not reflect the actual distances between the stimuli nor the actual duration of the stimuli.

Figure 2: Observed and predicted mean reaction times (RT) under 4 different spatial conditions are depicted as a function of SOA with auditory non-target leading. Note that RTs to visual stimulus decrease monotonically the earlier the auditory is presented. The points indicate observed RT including 95% confidence intervals: condition horizontal-coincident is presented by $\circ$; horizontal-disparate by $\Box$; vertical-coincident by $\triangleleft$; and vertical-disparate by $*$. Solid curves present the model predictions for horizontal, and dashed curves for vertical arrangements. The horizontal lines (blue) at the top mark observed RT to unimodal visual stimuli (solid: horizontal, dashed: vertical condition) and serve to gauge facilitation effects. Note that the displayed RT range is the same for all four participants.

Figure 3: Predictions of TWIN model for the probability of cuing and/or integration to occur within the same trial as function of SOA for participants S1-S4. Left column: horizontal configuration; right column: vertical configuration. Dotted (black) curve: integration and no cuing $[P(I \cap W^c)]$; dot-dashed (blue) curve: no integration but cuing $[P(I^c \cap W)]$; dashed (red) curve: integration and cuing $[P(I \cap W)]$; solid (cyan) curve: neither integration nor cuing $[P(I^c \cap W^c)]$. (i) probability of cuing without integration (dot-dashed/blue curve) decreases the later the auditory is presented; (ii) probability of both integration and cuing to occur (dashed/red curve) is maximal

at some mid-range SOA (−150 to −200 ms); (iii) probability of neither cuing nor integration to occur (solid/cyan curve) increases the later the auditory is presented (see text for further explication).

	S1			S2			S3			S4		
	M1	M2	M3	M1	M2	M3	M1	M2	M3	M1	M2	M3
$1/\lambda_A$	300	300	300	300	300	300	207	300	279	286	300	300
$1/\lambda_V$	19	117	5	33	105	29	38	158	5	22	146	5
μ	387	316	397	304	253	306	318	256	329	337	256	347
ω	250	250	–	250	190	–	250	250	–	250	250	–
ω_h	–	–	250	–	–	250	–	–	239	–	–	250
ω_v	–	–	185	–	–	243	–	–	77	–	–	184
Δ_{hc}	200	–	–	196	–	–	200	–	–	200	–	–
Δ_{hd}	164	–	–	144	–	–	146	–	–	153	–	–
Δ_{vc}	140	–	–	170	–	–	111	–	–	131	–	–
Δ_{vd}	84	–	–	127	–	–	63	–	–	100	–	–
Δ_c	–	46	179	–	136	183	–	200	200	–	81	174
Δ_d	–	0	130	–	77	134	–	132	126	–	42	130
Δ_w	–	165	–	–	138	–	–	143	–	–	160	–
α_h	–	0	–	–	77	–	–	0	–	–	0	–
α_v	–	109	–	–	111	–	–	206	–	-	111	–
k	8	9	7	8	9	7	8	9	7	8	9	7
χ^2	74	30	172	56	28	87	132	38	481	68	31	152
AIC	114	90	194	96	88	109	172	98	503	108	91	174

Table 1

Estimated parameters for the three different models (M1, M2, M3) individually for each participant (S1, S2, S3 ,S4). The last but one row contains χ^2 values; the last row, lists AIC values for goodness of fit and model comparison. k indicates the number of parameters.

Figure 1

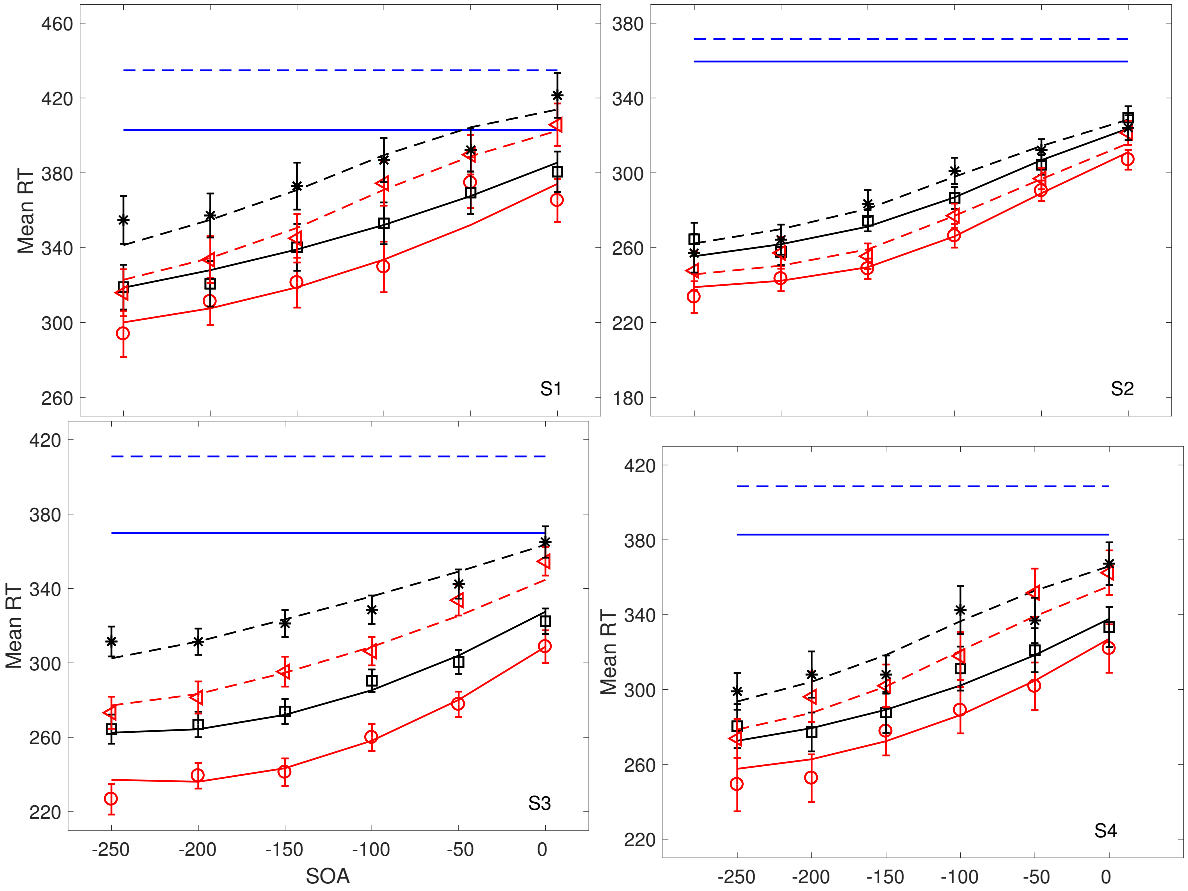
Mean RT
460
420
380
340
300
260
S1
380
340
300
260
220
180
S2
420
380
340
300
260
220
S3
-250
-200
-150
-100
-50
0
SOA
420
380
340
300
260
220
S4
-250
-200
-150
-100
-50
0

Figure 2

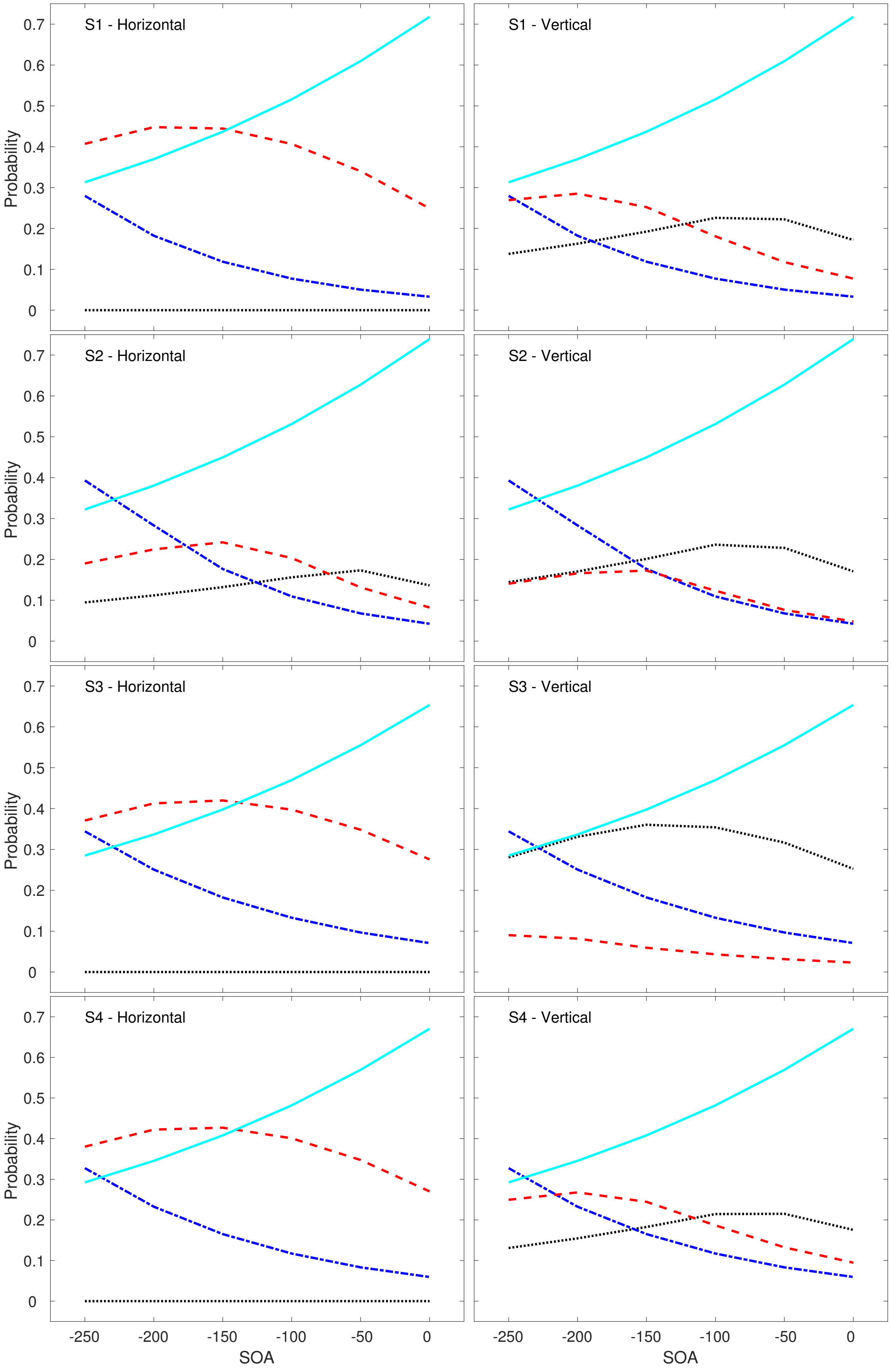
S1 - Horizontal
S1 - Vertical
S2 - Horizontal
S2 - Vertical
S3 - Horizontal
S3 - Vertical
S4 - Horizontal
S4 - Vertical
Probability
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
SOA
-250
-200
-150
-100
-50
0

Figure 3