

# The effect of spatial–temporal audiovisual disparities on saccades in a complex scene

Marc M. Van Wanrooij · Andrew H. Bell ·  
Douglas P. Munoz · A. John Van Opstal

Received: 1 October 2008 / Accepted: 11 April 2009 / Published online: 5 May 2009  
© The Author(s) 2009. This article is published with open access at Springerlink.com

**Abstract** In a previous study we quantified the effect of multisensory integration on the latency and accuracy of saccadic eye movements toward spatially aligned audiovisual (AV) stimuli within a rich AV-background (Corneil et al. in *J Neurophysiol* 88:438–454, 2002). In those experiments both stimulus modalities belonged to the same object, and subjects were instructed to foveate that source, irrespective of modality. Under natural conditions, however, subjects have no prior knowledge as to whether visual and auditory events originated from the same, or from different objects in space and time. In the present experiments we included these possibilities by introducing various spatial and temporal disparities between the visual and auditory events within the AV-background. Subjects had to orient fast and accurately to the *visual* target, thereby ignoring the auditory distractor. We show that this task belies a dichotomy, as it was quite difficult to produce *fast* responses (<250 ms) that were not *aurally* driven. Subjects therefore made many erroneous saccades. Interestingly, for the spatially aligned events the inability to ignore auditory stimuli produced shorter reaction times, but also more accurate responses than for the unisensory target conditions. These findings, which demonstrate effective multisensory integration, are similar to the previous study, and the same multisensory integration rules are applied

(Corneil et al. in *J Neurophysiol* 88:438–454, 2002). In contrast, with increasing spatial disparity, integration gradually broke down, as the subjects' responses became bistable: saccades were directed either to the auditory (fast responses), or to the visual stimulus (late responses). Interestingly, also in this case responses were faster and more accurate than to the respective unisensory stimuli.

**Keywords** Multisensory integration · Human · Gaze control · Race model · Natural scene

## Introduction

Saccadic eye movements reorient the fovea fast and accurately to a peripheral target of interest. Much of the neurophysiological mechanisms underlying saccades (Findlay and Walker 1999; Munoz et al. 2000, for review) have been revealed by studies carried out under simplified conditions, in which a single visual target evokes a saccade in an otherwise dark and silent laboratory room.

However, under more natural conditions, potential targets may be masked by a noisy audiovisual (AV) background. The brain should then segregate these targets from the background, weed out the irrelevant distractors, determine the target coordinates in the appropriate reference frame, and prepare and initiate the saccade. This is a highly nontrivial task, and it is thought that efficient integration of multisensory inputs could optimize neural processing time and response accuracy (Stein and Meredith 1993; Anastasio et al. 2000; Calvert et al. 2004; Colonius and Diederich 2004a; Binda et al. 2007).

Indeed, many studies have shown that combined auditory and visual stimuli lead to a significant reduction of saccade reaction times (SRTs; Frens et al. 1995; Hughes

---

M. M. Van Wanrooij · A. J. Van Opstal (✉)  
Department of Biophysics, Donders Institute of Brain,  
Cognition and Behaviour, Radboud University Nijmegen,  
Geert Grooteplein 21, 6525 EZ Nijmegen, The Netherlands  
e-mail: j.vanopstal@donders.ru.nl

A. H. Bell · D. P. Munoz  
Department of Physiology, Centre for Neuroscience Studies,  
CIHR Group in Sensory-Motor Systems, Queen's University,  
Kingston, ON K7L 3N6, Canada

et al. 1998; Colonius and Diederich 2004b). Theoretical analyses have shown that this reduction cannot be explained by mere statistical facilitation, an idea that is formalized by the so-called ‘race model’ (Raab 1962; Gielen et al. 1983). This principle holds that sensory inputs are engaged in a race, whereby the saccade is triggered by the sensory event that first crosses a threshold. This benchmark model predicts that, in the absence of any bimodal integration, the expected distribution of minimum reaction times shifts toward shorter latencies than those for the unimodal responses. Saccades elicited by simple AV stimuli show a general reduction of the SRT, in combination with a systematic modulation by the spatial–temporal stimulus separation (Frens et al. 1995; Hughes et al. 1998). Whereas the former effect may be attributed to statistical facilitation or to a nonspecific warning effect, the spatial–temporal modulation cannot be accounted for by the race model, and is a clear indication of AV integration. Spatial–temporal effects may be understood from neural interactions within a topographically organized multisensory representation. For that reason, the midbrain Superior Colliculus (SC) has been considered as a prime candidate for multisensory integration (Stein and Meredith 1993, for review; see also Anastasio et al. 2000, for theoretical accounts). Electrophysiological studies in the intermediate and deep layers of the SC have indicated that similar spatial–temporal interactions are found in the sensory and motor responses of saccade-related neurons (Meredith and Stein 1986; Meredith et al. 1987; Peck 1996, in cat; Frens and Van Opstal 1998; Bell et al. 2005, in monkey).

Yet, the majority of AV integration studies have typically been confined to the situation of one visual target, combined with one auditory stimulus (the latter often a distractor: Frens and Van Opstal 1995; Corneil and Munoz 1996; Harrington and Peck 1998; Colonius and Diederich 2004b). Few studies have quantified the effects of multisensory integration in more complex environments. In a recent study we investigated saccades to visual (V), auditory (A), and AV-targets in the two-dimensional frontal hemifield within a rich AV-background that contained many visual distractors and spatially diffuse auditory white noise (Corneil et al. 2002). The target could be a dim red LED, or a broadband buzzer. We systematically varied the signal-to-noise ratio (SNR) of the target sound versus background noise, to assess unisensory sound-localization behavior, and constructed spatially aligned AV-targets from the four different SNR’s and three onset asynchronies (12 different AV target types). In such a rich environment, the first V-saccade responses in a trial typically had long-reaction times, and were often in the wrong direction. A-saccades were typically faster than V-saccades, where the SNR primarily affected accuracy in stimulus elevation and saccade reaction time. Interestingly, all AV stimuli

manifested AV integration that could best be described by a “best of both worlds” principle: *auditory speed at visual accuracy* (Corneil et al. 2002, Fig. 10).

Note, that the subject’s task in these experiments was unambiguous: make a fast and accurate saccade to the target that appears as soon as the fixation light is extinguished. Yet, in more natural situations one cannot assume in advance that given visual and auditory events arose from the same object in space. In particular, as sound-localization is often less accurate than vision, perceived stimulus locations need not be aligned either. This was also the case in the experiments of Corneil et al. (2002), especially for the low SNR’s. However, the effect of *perceived* spatial misalignment (Steenken et al. 2008) was not investigated in that study.

Here, we describe AV integration for the situation that the subject has no advance knowledge about the spatial configuration of the stimuli. We thus extended the paradigm of Corneil et al. (2002) by introducing a range of spatial disparities between auditory and visual stimuli, and instructed the subject to localize the visual stimulus fast and accurately, and to ignore the auditory distractor. We varied the spatial and temporal disparities of AV stimuli, as well as the SNR of the auditory distractor against the background noise.

Although in the current experiments the auditory stimulus did not provide a consistent spatial cue for the visual target, we found that the saccadic system still efficiently used acoustic information to generate faster and more accurate responses for spatially aligned stimuli (presented in only 16% of the trials). We also obtained a consistent relation of the subject’s error rate with SRT for all (aligned and disparate) stimuli: for short SRT’s, saccades were acoustically guided, thus often ending at a wrong location. Late saccades were typically visually guided. For intermediate SRT’s to spatially disparate stimuli responses could either be auditorily or visually guided, but responses were still faster and more accurate than in the unisensory conditions. Similar bistable behavior has been reported for auditory and visual double-stimulation experiments engaged in target/non-target discrimination tasks (e.g., Ottes et al. 1985; Corneil and Munoz 1996). A theoretical account for our results is discussed.

## Methods

### Subjects

Five subjects, aged 24–44 (mean 30.2 years) participated in this study after having given their informed consent. All procedures were in accordance with the local ethics committee of the Radboud University Nijmegen. Three subjects

(A. John Van Opstal, JO; Andrew H. Bell, AB; and Marc M. Van Wanrooij, MW) are authors of this article; the remaining two (JG and JV) were naïve about the purpose of the study. Subjects JO and MW also participated in a similar previous study (Corneil et al. 2002). All subjects reported normal hearing and, with corrective glasses or lenses worn in the experimental setup (JG and JV), had normal binocular vision, except for JO, who is amblyopic in his right (recorded) eye. The eye signal calibration procedure (see below) was corrected for any nonlinearity that may have been present in this subject's data.

### Experimental setup

A detailed description of the experimental setup can be found in Corneil et al. (2002). Briefly, experiments took place in a completely dark and sound-attenuated room, in which echoes above 500 Hz were effectively attenuated and the overall background sound level was about 30 dB, A-weighted (dBA). Subjects were seated facing a rich stimulus array with their head supported by an adjustable neck rest. Horizontal and vertical eye movements were recorded using the scleral search coil technique (Robinson 1963; Collewyn et al. 1975), sampled at 500 Hz/channel.

The stimulus array consisted of 85 light emitting diodes (LED) mounted onto a thin wire frame at 85 cm in front of the subject (Fig. 1). The LEDs were arranged in 7 concentric circles at eccentricities  $R \in [2; 5; 9; 14; 20; 27; 35]^\circ$ , and placed at 12 different directions ( $\Phi \in [0; 30; 60; \dots; 330]^\circ$ , where  $\Phi = 0^\circ$  is rightward,  $\Phi = 90^\circ$  is upward, etc. (Fig. 1). All LEDs could illuminate either red (0.18 cd/m<sup>2</sup>) or green (0.24 cd/m<sup>2</sup>). To produce the visual background all 85 LEDs were turned green. The visual fixation point at  $[R, \Phi] = [0, 0]^\circ$  and the target was subsequently specified by turning the appropriate LED from green to red.

The auditory background was generated by a circular array of nine speakers (Nellcor), mounted onto the wire frame at about 45° eccentricity (Fig. 1a). Sound intensities were measured at the position of the subject's head with a calibrated sound amplifier and microphone (Brüel & Kjaer BK2610/BK4144, Norcross, GA), and are expressed in dBA. The auditory background consisted of broadband Gaussian white noise (0.2–20 kHz) at a fixed intensity of 60 dBA. The auditory distractor stimulus was produced by a broadband lightweight speaker (Philips AD-44725, Eindhoven, the Netherlands) mounted on a two-link robot, which allowed the speaker to be positioned in any direction at a distance of 90 cm (Hofman and Van Opstal 1998). The auditory distractor stimulus consisted of a periodic broad-band noise (period 20 ms, sounding like a 50 Hz buzzer) that had a flat broad-band characteristic between 0.2 and 20 kHz, presented at a variable intensity (see below).

### Paradigms

Subjects completed three different paradigms: a visual calibration paradigm, an auditory localization paradigm, and the AV distractor paradigm that contributed to the bulk of the experimental data. Every session began with the visual calibration paradigm followed by 2–4 blocks of the auditory localization and/or AV distractor paradigms.

#### *Visual calibration*

Subjects were required to generate saccades to visual stimuli pseudo-randomly presented to 1 of 60–72 possible target locations (12 directions, 5–6 different eccentricities between 5 and 35°) in the absence of the AV-background. Each trial began by turning the central LED red (fixation point) for 800 ms. When it extinguished, a peripheral red target LED was illuminated which the subject had to re-fixate. Each target location was presented once. Similar to Corneil et al. (2002), the final saccadic endpoint was used for calibration purposes, whereas the endpoint of the first saccade was used for the visual-only data ( $V_{\text{NOBG}}$ , without background).

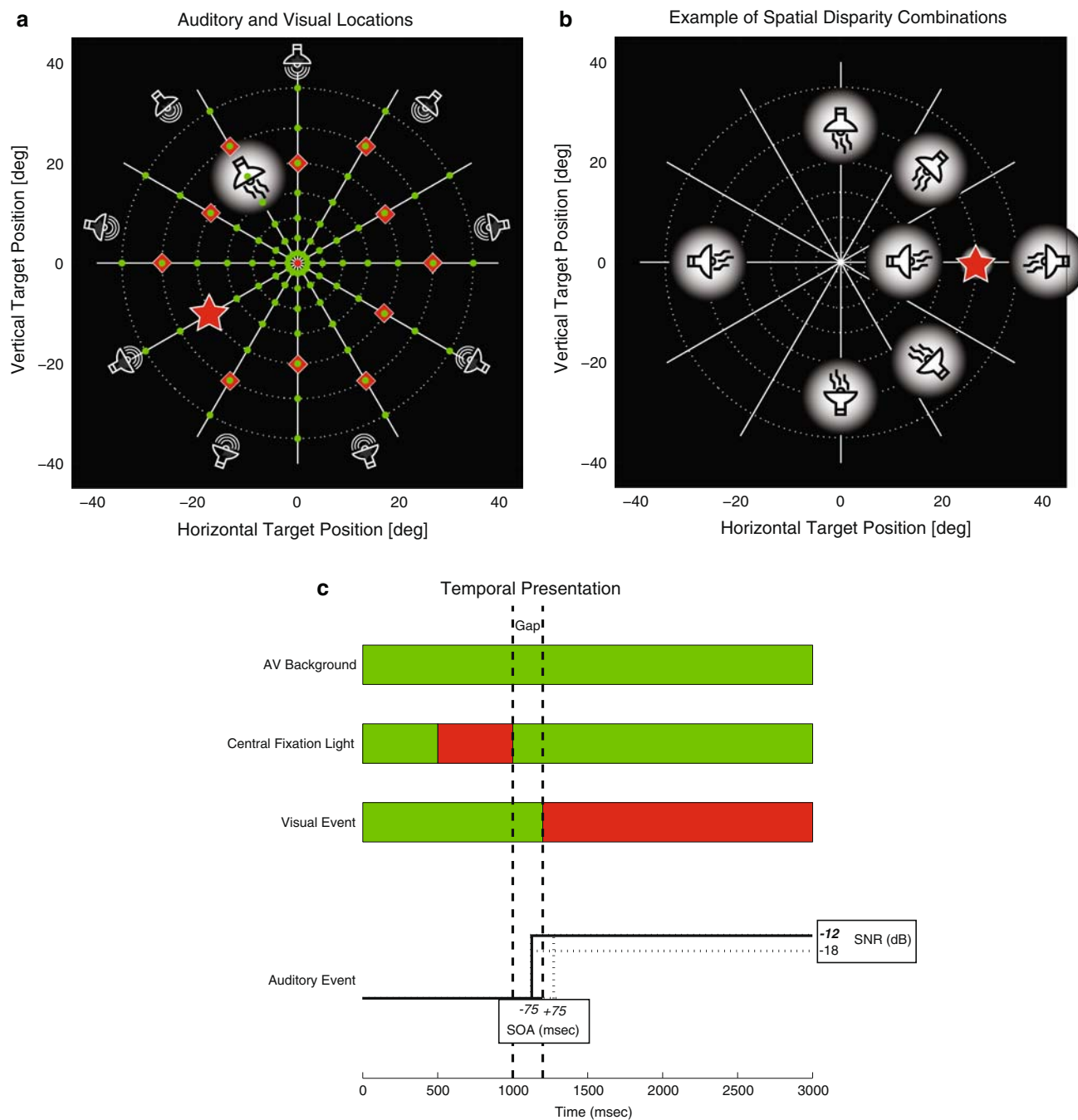
#### *Auditory localization*

Subjects generated saccades to auditory targets in the presence and absence of the AV-background (A and  $A_{\text{NOBG}}$ , respectively). These data served to assess sound-localization performance under different SNR conditions. Each trial began with fixation of the central visual fixation point for 600–850 ms. Then, an auditory target was presented from 1 out of 25 possible locations within the oculomotor field. Auditory targets were presented at four different SNRs relative to the acoustic background (−6, −12, −18, −21 dB). A- and  $A_{\text{NOBG}}$ -trials were run in separate blocks, often within the same experimental session.

#### *Audiovisual distractor paradigm*

Subjects generated saccades amidst an AV-background to V- and AV-targets. Each trial began with the appearance of the AV-background (Fig. 1a). After a randomly selected delay of either 150, 275, or 400 ms, the central LED turned red, which the subject had to fixate for 600–850 ms. The fixation LED was then turned green, and after a 200 ms gap a peripheral red target LED was illuminated. Subjects had to generate a saccade quickly and accurately to the peripheral target LED. The location of the target was selected pseudo-randomly from 1 out of 12 possible locations (12 directions,  $R = 20, 27^\circ$ ; Fig. 1a).

An auditory distractor was chosen from 8 possible locations for the visual target (Fig. 1b): aligned (distractor



**Fig. 1** Audiovisual paradigm. **a** Spatial representation of stimulus events. Subjects had to make a saccade to a red visual (*red star*) stimulus, ignoring an auditory distractor (*highlighted speaker*). The background consisted of diffuse noise from 9 small speakers and 85 green LEDs (*green dots*). Visual targets were randomly chosen from 12 locations (*red diamonds*). **b** The position of the distractor could

coincide with the visual target (*red star*), or deviate in either direction ( $\pm 45$ ,  $\pm 90$ ,  $180^\circ$ ), or eccentricity (by a factor of 0.67 or 1.4–1.6). **c** Temporal events in a trial. The distractor could be presented at one of two temporal asynchronies ( $\pm 75$  ms), and one of two SNRs ( $-12$  and  $-18$  dB for background)

presented behind the target LED); same direction but either 0.67 or 1.4–1.6 times the radial eccentricity; and same eccentricity but rotated at  $\pm 45$ ,  $\pm 90$ , or  $180^\circ$  away from the target location. The visual target either preceded the distractor by 75 ms (V75A), or followed by 75 ms (A75V)

at equal probability. The auditory distractor was presented at one of two possible SNRs:  $-12$  dB ( $A_{12}$ ) or  $-18$  dB ( $A_{18}$ ) relative to the background (Fig. 1c). We also interspersed 36 V-only trials (selected from 12 possible directions and three eccentricities: 14, 20, and  $27^\circ$ ). Thus, the

total number of different trials in the AV paradigm was 420 ([12 target locations × 8 spatial disparities × 2 asynchronies × 2 SNRs] + 36 V conditions). These were presented as four blocks of 105 randomly selected trials, with a short rest in between. The order of the blocks was randomized from session to session; each subject completed multiple blocks yielding 1004–2391 saccades per subject (Table 1). Note that in each block only 15–18% of all trials contained spatially aligned AV stimuli.

Data analysis

All data analysis was performed in MatLab 7.4 (The Matworks, Inc.).

Data calibration

Response data were calibrated by training two three-layer neural networks with the back-propagation algorithm that mapped final eye positions onto the target positions of the visual calibration paradigm (Goossens and Van Opstal 1997). Eye-position data from the other paradigms were then processed using these networks, yielding an absolute accuracy <3% over the entire range. Saccades were automatically detected from calibrated data, based on velocity and acceleration criteria using a custom-made program. Onset and offset markings were visually checked by the experimenter, and adjusted if necessary.

Coordinate systems

Target and response coordinates are expressed in the azimuth and elevation coordinates of the double-pole

coordinate system (Knudsen and Konishi 1979), and is related to the spherical polar angles ( $R, \varphi$ ) of the LEDs:

$$\alpha = \sin^{-1}(\sin R \cos \varphi) \text{ and } \varepsilon = \sin^{-1}(\sin R \sin \varphi). \quad (1)$$

with  $\alpha, \varepsilon, R,$  and  $\varphi$  azimuth, elevation, eccentricity, and direction, respectively. The inverse relations read:

$$\begin{aligned} \varphi &= \tan^{-1}(\sin \varepsilon / \sin \alpha) \text{ and } R \\ &= \tan^{-1}\left(\sqrt{\sin^2 \alpha + \sin^2 \varepsilon} / \sqrt{\cos^2 \alpha - \sin^2 \varepsilon}\right). \end{aligned} \quad (2)$$

Reaction times

Race model

We compared the observed multisensory SRT distributions to the race model (Raab 1962):

$$P(\tau_{AV} \leq t) = P(\tau_A \leq t) + P(\tau_V \leq t) - P(\tau_A \leq t, \tau_V \leq t) \quad (3)$$

with  $P(\tau \leq t)$  the probability of observing a reaction time  $\tau$  that is faster than or equal to a specified time  $t$ . There is actually a whole class of race models, each corresponding to a different joint distribution for  $\tau_A$  and  $\tau_V$ . They all satisfy the following two distribution inequalities (Colonius 1990):

$$P(\tau_{AV} \leq t) \leq \min(P(\tau_A \leq t) + P(\tau_V \leq t), 1) \quad (4)$$

and

$$\max(P(\tau_A \leq t), P(\tau_V \leq t)) \leq P(\tau_{AV} \leq t). \quad (5)$$

Both positive violations of Eq. 4 (upper bound, corresponding maximally to negative-dependency) and negative violations of Eq. 5 (lower bound) indicate

**Table 1** Stimulus types used in the experiments

Stimulus types						Number of responses					
Abbr	Target	Distractor	SNR	SOA	BCKGR	MW	JG	JO	AB	JV	Total
V <sub>NOBG</sub>	Visual	–	–	–	–	786	310	652	684	359	2791
V	Visual	–	–	–	+	135	76	139	125	43	518
A <sub>18,NOBG</sub>	Auditory	–	–18	–	–	75	0	50	61	50	236
A <sub>21,NOBG</sub>	Auditory	–	–21	–	–	75	0	50	66	50	241
A <sub>18</sub>	Auditory	–	–18	–	+	74	18	75	50	50	267
A <sub>21</sub>	Auditory	–	–21	–	+	73	15	75	50	47	260
V75A <sub>18</sub>	Visual	Auditory	–18	+75	+	306	181	296	259	108	1,150
V75A <sub>21</sub>	Visual	Auditory	–21	+75	+	288	180	290	253	99	1,110
A <sub>18</sub> 75V	Visual	Auditory	–18	–75	+	287	171	289	251	97	1,095
A <sub>21</sub> 75V	Visual	Auditory	–21	–75	+	292	170	295	250	101	1,108
Total						2,391	1,121	2,211	2,049	1,004	8,776

Abbr: Abbreviations for the various stimulus types. Target: Target modality (either auditory or visual). Distractor: present or absent. SNR: The signal-to-noise ratio of the auditory distractor. SOA: stimulus onset asynchrony, either –75 ms (auditory leading) or +75 ms (visual leading). BCKGR: presence of the AV-background



convergence of auditory and visual inputs and may provide evidence for multisensory integration. We have tested (*t*-test, significance level of  $P = 0.05$ ) for violations of both inequalities (Colonius and Diederich 2006; Ulrich et al. 2007), and made a comparison to a special case of the race model:

$$P(\tau_{\text{race}} < t) = P(\tau_A < t) + P(\tau_V < t) - P(\tau_A < t) \times P(\tau_V < t). \quad (6)$$

This model implies stochastic independence between the two channels (Meijers and Eijkman 1977; Gielen et al. 1983), and was also used in our previous study (Corneil et al. 2002).

### Bistable behavior

As will become clear in the Results, subjects often displayed bistable localization responses, in which they appeared to make a saccade either to the auditory distractor, or to the visual target. None of the race models can account for such bimodal<sup>1</sup> behavior. The simplest version of a bistable mechanism assumes “either A or V” response behavior, and could be formulated as:

$$P(\tau_B) = \alpha P(\tau_A) + (1 - \alpha)P(\tau_V) \quad (7)$$

With  $P(\tau_B)$  the predicted bistable distribution,  $\alpha$  the probability of responding to an auditory stimulus, and  $1-\alpha$  the probability of responding to a visual stimulus, the bistable SRT distributions will resemble the weighted summed distributions of the A and V- saccades in which the probability  $\alpha$  acts as weighting parameter. Once again, as with the race model, deviations from this independent model may indicate multisensory integration. Figure 2b and c illustrates the predictions of the stochastically independent (Eq. 6) race and bistable models for simulated data (Fig. 2a).

### Localization accuracy

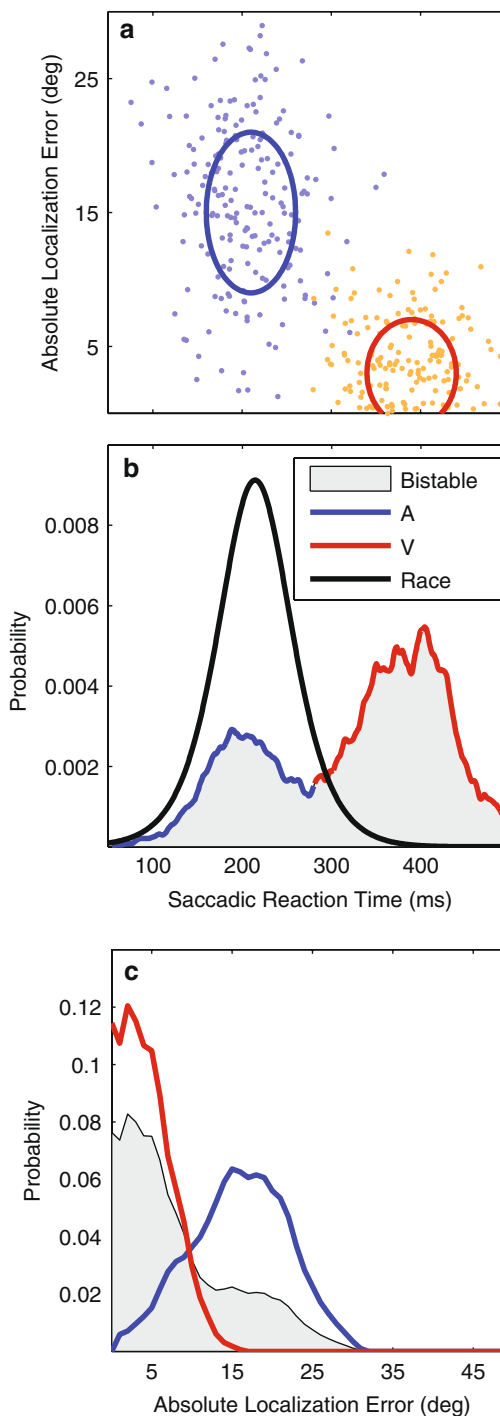
#### Regression

We quantified localization behavior by linear regression on the stimulus–response relation:

$$\alpha_R = a \alpha_T + b \text{ and } \varepsilon_R = c \varepsilon_T + d. \quad (8)$$

with  $\alpha_R$ ,  $\alpha_T$ ,  $\varepsilon_R$ , and  $\varepsilon_T$  response azimuth, target azimuth, response elevation, and target elevation, respectively. Parameters  $[a, b, c, d]$  were found by minimizing the mean squared error (Press et al. 1992). From this fit we also

<sup>1</sup> We use the terms “unimodal” and “bimodal” in a statistical sense (single- and double-peaked distributions, respectively), without referring to the unisensory or multisensory origin of the response distributions.



**Fig. 2** Predictions of models. **a** Hypothetical distributions of SRT-localization errors for unisensory responses (blue: A-only, orange: V-only). Blue and red ellipses denote 1 SD around the mean. A—responses are fast but inaccurate; V—saccades are accurate but slow. **b** Distribution of SRTs according to two conceptual models: gray shading: bistable model (with  $\alpha = 1/3$ , Eq. 7), black curve: race model (Eq. 6). The blue–red curve indicates a higher probability of eliciting an auditory response in blue and a visual response in red. **c** Distribution of AV saccade errors according to the bistable model (gray shading). Blue and red curves indicate the unisensory error distributions (auditory and visual, respectively). Note that the race model would predict an error distribution close to the auditory response distributions

determined the correlation coefficient between data and fit, and the mean absolute error of the responses.

*Modality index and perceptual disparity*

Because the perceived location of a stimulus (evidenced by the response) does not necessarily coincide with its physical location, the spatial disparities of the AV stimuli should be defined appropriately. The physical AV stimulus disparity, SD, is:

$$SD = \sqrt{(\alpha_{T,V} - \alpha_{T,A})^2 + (\varepsilon_{T,V} - \varepsilon_{T,A})^2} \tag{9}$$

The perceptual AV-disparity was then defined by the responses to unisensory auditory and visual targets for the given condition:

$$PD = \sqrt{(\alpha_{R,V} - \alpha_{R,A})^2 + (\varepsilon_{R,V} - \varepsilon_{R,A})^2} \tag{10}$$

where  $\alpha_R$  and  $\varepsilon_R$  were obtained from the linear fits on A and V stimulus–response data (Eq. 7). The perceptual localization error of an AV stimulus with regard to the respective unisensory percepts was then determined by:

$$PE_V = \sqrt{(\alpha_{R,V} - \alpha_{R,AV})^2 + (\varepsilon_{R,V} - \varepsilon_{R,AV})^2} \text{ and } PE_A = \sqrt{(\alpha_{R,A} - \alpha_{R,AV})^2 + (\varepsilon_{R,A} - \varepsilon_{R,AV})^2}. \tag{11}$$

These measures quantify the distance (in deg) between an AV response and the perceived unisensory locations of V and A, respectively. Finally, from the perceptual localization errors we defined a dimensionless modality index, MI:

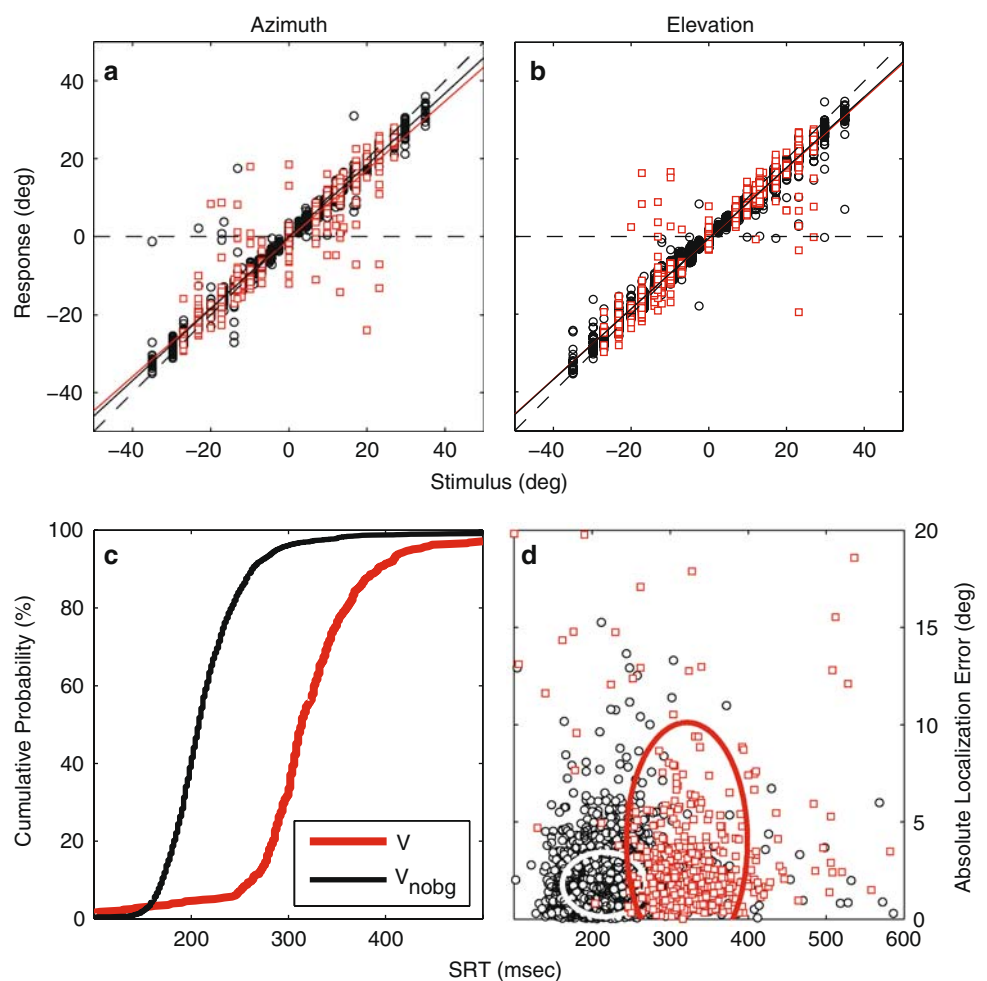
$$MI = \frac{PE_V - PE_A}{PE_V + PE_A} \tag{12}$$

which indicates, for each AV response, whether it lies closer to the A percept (MI = +1, as  $PE_V \gg PE_A$ ) or V percept (MI = -1;  $PE_A \gg PE_V$ ). A value of MI  $\approx$  0 suggests an integrated AV percept (see Fig. 8).

*Overview*

In the analyses presented here, we pooled data across subjects, unless noted otherwise. Statistical significance of a difference between two distributions was assessed by the 1D or 2D KS-test, where we took  $P < 0.05$  as the accepted

**Fig. 3** Effect of AV-background on V-saccades. Red/black symbols: V/V<sub>NOBG</sub>. **a, b** Stimulus–response plots of endpoints of primary V-saccades against stimulus **a** azimuth and **b** elevation. **c** Cumulative SRT probability functions. **d** Absolute localization error as a function of SRT. Ellipses circumscribe 2 SD around the mean. Data pooled across subjects



level of significance. The analysis was thus based on a total of 8776 trials. Table 1 gives a detailed breakdown of trials per subject.

## Results

We first quantify the basic properties of the V- and A-saccades in our experiments, as they are crucial for later comparisons with the AV-responses.

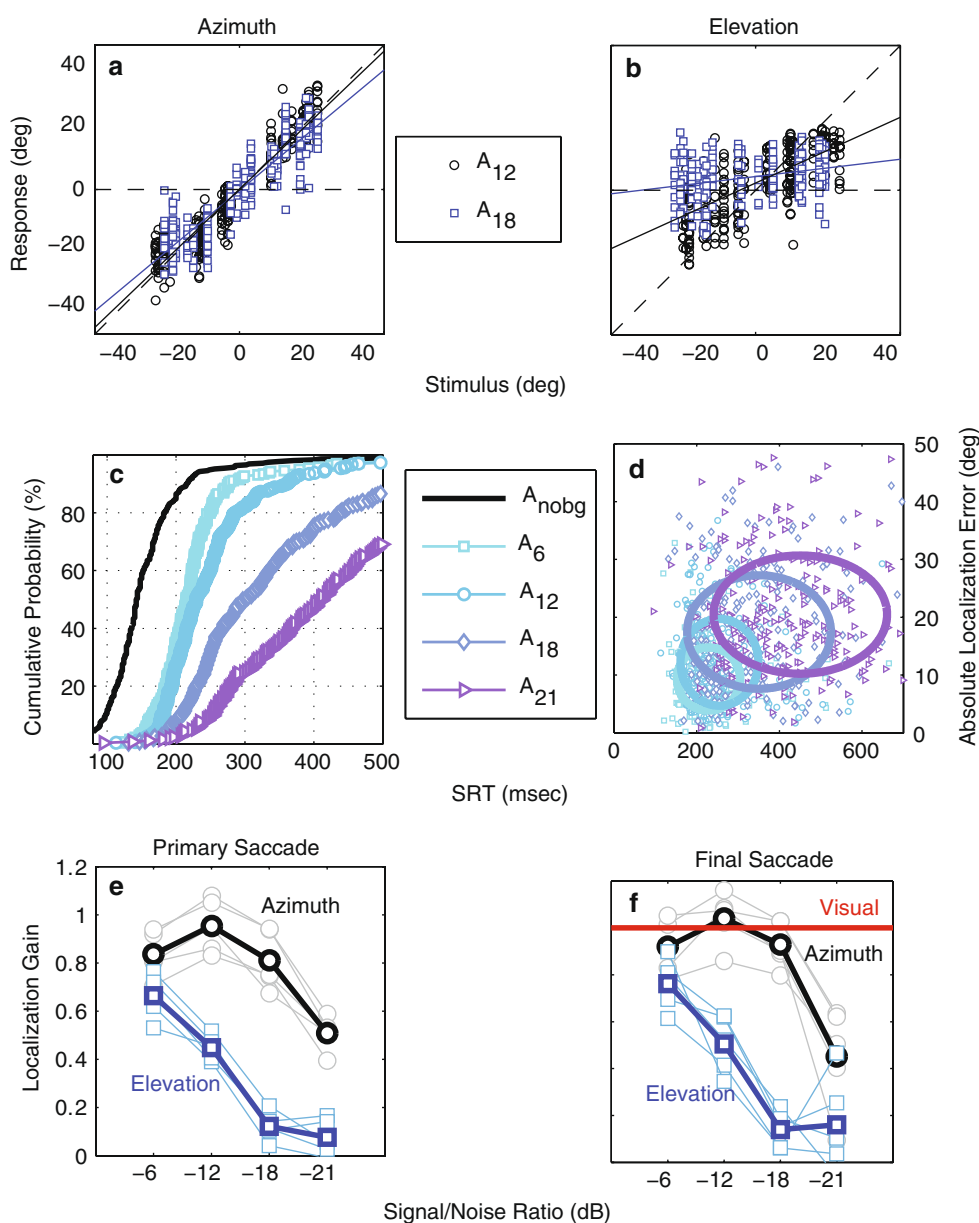
### V- and A-saccades

The AV-background hampered localization accuracy of unisensory visual targets (Fig. 3a, b). V-trials displayed a

larger amount of scatter in primary saccade responses than visual trials in the no-background condition ( $V_{\text{NOBG}}$ , Fig. 3a, b: red squares and black circles, respectively), both in azimuth (Fig. 3a) and in elevation (Fig. 3b). This resulted in lower correlations between stimulus and response ( $r^2 = 0.98$  for  $V_{\text{NOBG}}$  and  $\sim 0.89$  for V,  $P \ll 0.001$ ). The subject's SRT increased by about 100 ms in the presence of the AV-background (Fig. 3c). The 2D distributions of absolute localization error versus SRT for both conditions (Fig. 3d;  $V_{\text{NOBG}}$  and V: black circles and red squares, respectively) are clearly distinguishable from one another (2D KS-test,  $P \ll 0.001$ ).

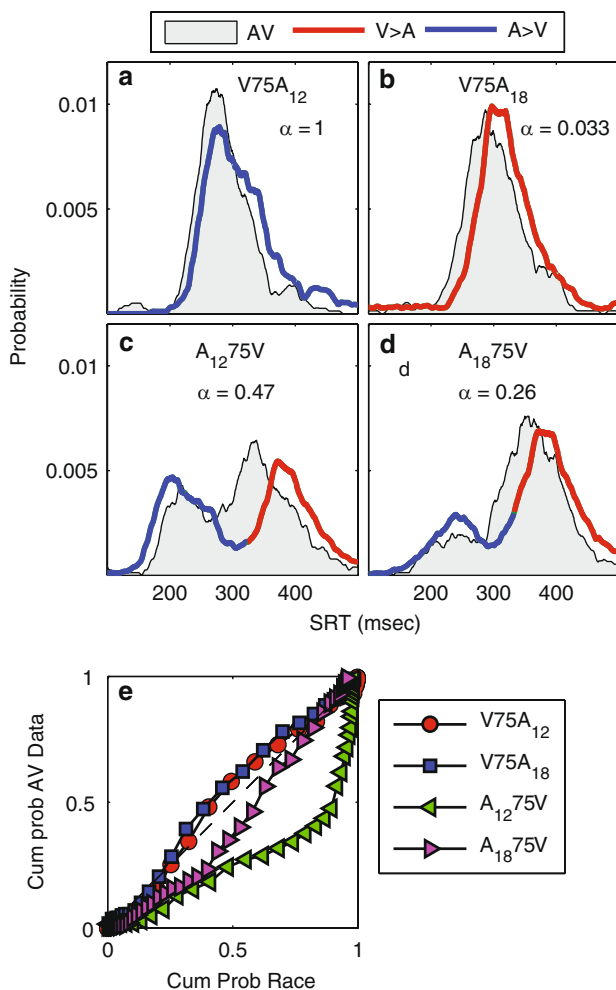
Localization performance of A-saccades was compromised even more by the AV-background albeit in different ways than V-saccades. First, the azimuth and elevation

**Fig. 4** Effect of AV-background of A-saccades. **a, b** Stimulus–response plots of A-saccade endpoints against **a** stimulus azimuth and **b** elevation. *Blue/black symbols: A/A<sub>NOBG</sub>.* **c** Cumulative SRT probability functions for the A<sub>NOBG</sub>- (*black line*), A<sub>6</sub>-, A<sub>12</sub>-, A<sub>18</sub>-, and A<sub>21</sub> saccades. **d** Absolute localization error plotted as a function of SRT. Data pooled across subjects. *Ellipses circumscribe 2 SD around the mean.* **e** Response gains for azimuth (*gray circles; black: mean across subjects*) and elevation (*cyan squares; blue: mean across subjects*) of primary A-saccades as a function of SNR. **f** Final saccade response gains for V- (*red*) and A-saccades





components of A-responses were affected differently (Fig. 4a, b). For example, the  $A_{12}$  responses (Fig. 4a, b, black circles) were more accurate in azimuth than in elevation (i.e., less scatter and a higher response gain). This property results from the different neural processing pathways of sound–location coordinates (binaural difference cues, for azimuth, versus pinna-related spectral shape cues, for elevation; e.g., Oldfield and Parker 1984; Blauert 1997; Hofman and Van Opstal 1998). Second, localization performance depended strongly on the SNR too. The  $A_{18}$  responses (Fig. 4a, b, purple squares) had a lower gain and more scatter in elevation than the  $A_{12}$  responses. Furthermore, the SRTs were prolonged for decreasing SNRs



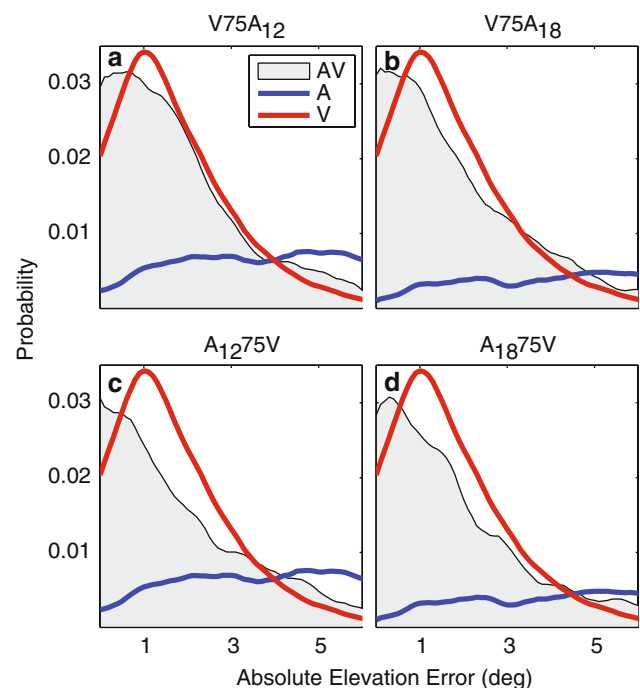
**Fig. 5** Comparison of SRT distributions for AV-aligned trials. **a–d** The bisensory SRT distributions (gray patch) for **a** V75A<sub>12</sub>, **b** V75A<sub>18</sub>, **c** A<sub>12</sub>/75V, and **d** A<sub>18</sub>/75V, are shown as probability distributions. **a–d** Predicted bistable distributions (Eq. 7) are shown as red–blue curves, with blue indicating a larger probability of responding to A at a given SRT, and red indicating a larger probability of responding to V. Parameter  $\alpha$  is probability of responding to an auditory stimulus in a given trial (irrespective of SRT, Eq. 7). **e** Comparison of bisensory SRTs to the predictions of the stochastic-independent race model (Eq. 6)

(CDFs in Fig. 4c). Also the distributions of absolute localization error versus SRT for the various SNRs clearly differed from one another (Fig. 4d). These features are summarized in Fig. 4e, which shows that azimuth gain (black circles) dropped for decreasing SNRs, but the elevation gain (blue squares) dropped even faster. These results are in accordance with earlier studies that reported a degrading effect of background noise on sound-localization performance (Good and Gilkey 1996; Zwiers et al. 2001; Corneil et al. 2002).

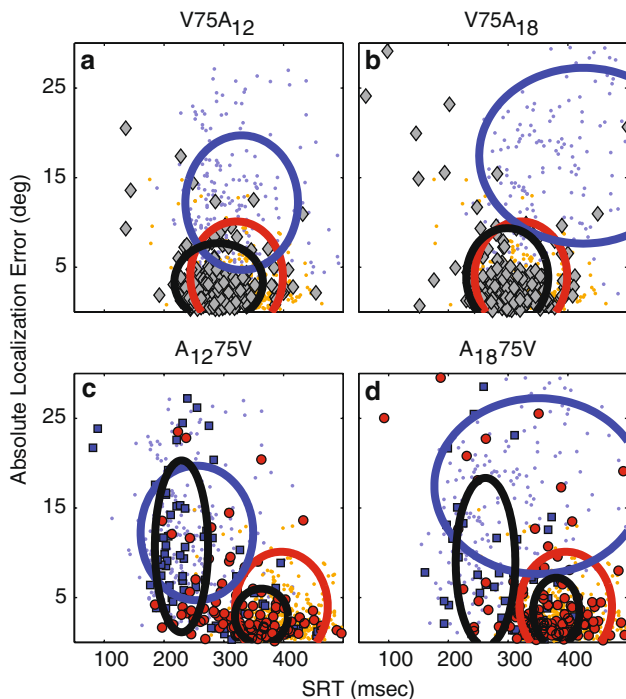
An important difference between V- and A-saccades, which cannot be readily observed from the primary saccade responses, is the difference in localization percepts induced by the AV-background (Fig. 4f). Although it could take a few attempts/saccades, subjects eventually localized the V-target (red line). In contrast, the background noise introduced a large undershoot in azimuth and elevation also for the final A-saccades. This aspect is important for the AV-disparity experiment, since the stimulus disparity between A- and V-targets deviated from the perceptual disparity. We will return to this difference in a later section.

#### Spatially aligned AV stimuli

In only 16% of the AV trials the auditory stimuli were spatially coincident with the visual target. Subjects were



**Fig. 6** Comparison of localization error distributions in elevation. **a–d** The bisensory (gray patch) and unisensory (V, red; A, blue) error distributions for **a** V75A<sub>12</sub>, **b** V75A<sub>18</sub>, **c** A<sub>12</sub>/75V, and **d** A<sub>18</sub>/75V. Note that, for each condition, the AV distribution is shifted toward smaller errors than the best unisensory (visual) distribution



**Fig. 7** Absolute localization error as a function of SRT. The 2D response distributions for unisensory visual (orange dots, red ellipse) and auditory targets (blue dots and ellipse) are shown in (a–d) for comparison with the four spatially aligned bisensory response distributions: **a** V75A<sub>12</sub> (gray diamonds), **b** V75A<sub>18</sub> (gray diamonds), **c** A<sub>12</sub>75V (blue squares and red circles), and **d** A<sub>18</sub>75V (blue squares and red circles). Either the unisensory visual or auditory distribution is shifted by  $\pm 75$  ms, to align SRTs with the first stimulus of the AV-trials. Ellipses circumscribe 1 SD around the mean. Only unisensory data within 2 SD of the mean are shown. The blue squares and red dots in (c) and (d) were obtained through K-means cluster analysis on the bisensory data (mean silhouette-values: 0.76 and 0.73, respectively). Black ellipses indicate means and SD of the clusters. Note that bisensory distributions have on average reduced SRTs and smaller localization errors than the unisensory distributions

asked to localize the visual target fast and accurately regardless the auditory distractor. Here, we first analyze responses to these stimuli, to check whether AV interactions would still follow the same rules as in the Corneil et al. (2002) study.

Figure 5a–d presents the SRT distributions for the unisensory (A, blue; V, red) and AV (gray patch) stimuli. The V75A stimuli (auditory lagging; Fig. 5a, b) both exhibit a single-peaked distribution with shorter SRTs than either unisensory distribution. This multisensory enhancement exceeds the prediction of statistical facilitation by the stochastically independent race model (Fig. 5e, Eq. 6). This held in particular for the V75A<sub>18</sub> stimulus, which also exceeded the negative-dependency race model (Eq. 4). This phenomenon of largest enhancement for weakest stimuli has been termed *inverse effectiveness* in the neurophysiological literature (Stein and Meredith 1993).

In contrast, the A75V stimuli (auditory leading; Fig. 5c, d) both produced bimodal SRT distributions, with longer SRTs than the fastest A-distribution. Interestingly, bimodal response distributions were not obtained in the Corneil et al. (2002) study (see also “Discussion”). Note that the stochastically independent race model (Eq. 6) is also violated for these stimuli (Fig. 5e), as it predicts a single-peaked, faster (or equally fast) SRT distribution for all AV stimuli (the response SRTs even fail to reach the lower bound of the race model of Eq. 5, not shown). Yet, the measured distribution does not coincide with the predicted bistable response distribution of Eq. 7 (e.g., Fig. 2b) either. Thus, we conclude that both AV stimulus types underwent multisensory integration.

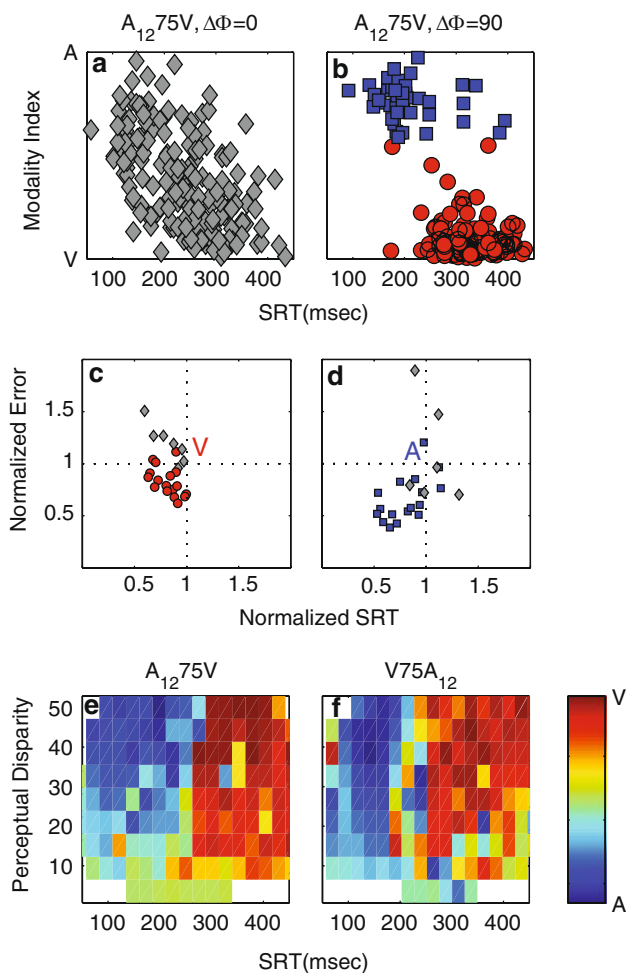
Corneil et al. (2002) also showed that localization errors for aligned AV stimuli were smaller than for either unisensory stimulus. Figure 6 demonstrates that this was also true in the distractor paradigm. For all four aligned AV conditions (Fig. 6, gray patch), subjects localized more accurately than the V condition (Fig. 6: V red; A blue).

To obtain an integrated overview of these data, Fig. 7a–d compares the response distributions of absolute localization error versus SRT. Note that the V75A stimuli (Fig. 7a, b) yielded a single-cluster of AV-responses (gray diamonds with black ellipse at 1 SD around the mean), with an average SRT and error that was smaller than either unisensory response distribution (A-saccades: small blue dots and ellipse; V-saccades: small yellow dots and red ellipse).

In contrast, two response clusters might be expected for A75V stimuli, corresponding to bistable responses (Fig. 5c, d). We therefore performed a K-means clustering analysis ( $K = 2$ , based on SRT, response azimuth, elevation, eccentricity, and direction), which indeed divided the data into distinct distributions (labeled by blue squares and red circles; Fig. 7c, d) with relatively high silhouette-values (0.76 for A<sub>12</sub>75V and 0.73 for A<sub>18</sub>75V).

The separated clusters (black ellipses) can be readily compared to the straightforward bistable model, which would yield two AV-clusters coinciding with either unisensory V- and A-distribution (Fig. 2). For the V75A stimuli and also for larger numbers of clusters on the A75V stimuli, the silhouette-values quickly dropped to values  $< 0.5$ , indicating that a larger number of clusters is not readily observed in the data.

Taking a coarser look at the A75V data (Fig. 7c, d), the blue cluster best resembles the A-distribution, while the red cluster resembles the V-distribution. Yet, some responses in both clusters have SRTs and errors that could have resulted from either cluster. A better look at the data reveals a *gradual* improvement in localization error as reaction time progresses, rather than a sudden drop that would have resulted from a true bistable mode as subjects would have shifted from fast and inaccurate auditory, to



**Fig. 8** Performance for spatially disparate conditions. **a** Modality index as a function of SRT for spatially aligned  $A_{12}75V$  stimuli. MI gradually runs from values near +1 (A-response) to  $-1$  (V-response) in this single-clustered dataset. **b** Disparate condition ( $\Delta\Phi = \pm 90^\circ$ ). K-means cluster analysis yielded two distinct clusters: *blue squares*, cluster 1 data; *red circles*, cluster 2 data. Cluster 1 represent aurally driven saccades, cluster 2 are visually triggered responses. **c** Error versus SRT for all 24 AV conditions of the cluster 2 data, normalized by mean V-saccade error and SRT (*red circles*) **d** Error versus SRT for all 24 AV conditions of the cluster 1 data, normalized by mean A-saccade error and SRT (*blue squares*). *Gray squares* in (**c**, **d**) represent conditions that yielded single-clustered data (because of low silhouette-values [ $<0.6$ ] or a small number of responses in one of the two K-means clusters [ $<5$ ]). **e** MI as a function of the perceived disparity (in deg) and SRT for  $A_{12}75V$  stimuli. **f** Same for  $V75A_{12}$  stimuli

slow but accurate visual responses. In fact, at any given SRT AV-responses were more accurate than the unisensory responses, which further underline the evidence for multisensory interaction.

#### Spatially disparate AV stimuli

Figure 8a, b shows the distributions of the AV modality index (“Methods”, Eq. 11) versus SRT for two AV

conditions. The MI is a measure for the resemblance of a particular AV response to either a unisensory V or A-saccade. Note that it is expressed in terms of the *perceived*, rather than the *physical* disparity, so that even the spatially aligned stimuli (e.g., Fig. 8a) can be shown to have evoked both aurally and visually driven saccades (MI close to +1 and  $-1$ , respectively). Interestingly, the spatially aligned  $A_{12}75V$  (Fig. 8a) data seemed to consist mostly of intermediate AV-responses, and MI gradually shifted from +1 to  $-1$  as time progressed. K-means clustering of these data on two or more clusters yielded low silhouette-values ( $<0.6$ ) and few responses ( $<6$ ) were assigned to one of the clusters.

For  $A_{12}75V$  stimuli with a considerable angular disparity (here  $\Delta\Phi = \pm 90^\circ$ ; Fig. 8b), however, K-means cluster analysis produced two clear distributions that appeared to obey the principles of a bistable mechanism: either auditory (blue), or visual (red) responses.

Figure 8c, d summarizes our findings for all 24 AV stimulus conditions employed in this study. In 17/24 conditions the response data could be separated into two clusters (single-cluster conditions:  $V75A_{12}$ ,  $\Delta\phi = 90$  and  $\Delta R = 1.5$ ;  $V75A_{18}$ ,  $\Delta\phi = 0$  and  $\Delta\phi = 180$ ;  $A_{12}75V$ ,  $\Delta\phi = 0$ ;  $A_{18}75V$   $\Delta\phi = 0$  and  $\Delta\phi = 90$ ). Figure 8c normalizes the cluster with the longest SRT against the V-responses, whereas in Fig. 8d the cluster with the shortest SRT was normalized against A-saccades ( $-12$  and  $-18$  dB). If these responses would follow the simple bistable model of Fig. 2, all points would scatter around the center of these plots. As data points lie predominantly in the lower-left quadrant, the interesting point of this analysis is, that for all stimulus conditions responses were actually *better* (i.e., faster and more accurate) than pure V- and A-saccades. Hence, even for spatially unaligned stimuli, AV enhancement occurs and the simple bistable model should be rejected.

Figure 8e, f summarizes our analysis for all perceived disparities of the  $A_{12}75V$  and  $V75A_{12}$  stimuli. A clear pattern emerges in this plot: only when *perceived* disparity is very small, MI is close to zero (green-colored bins), indicative for multisensory integration. It rapidly splits into two clusters for larger perceived disparities, with invariably aurally guided responses (blue) for the short SRTs ( $<250$  ms), and visually guided saccades for longer SRTs (red). Hence, these plots delineate a sharply-defined spatial-temporal window of AV integration. Similar results were obtained for the  $A_{18}$  distractor (not shown).

#### Discussion

We studied the responses of the human saccadic system when faced with a visual orienting task in a rich AV

environment and a competing auditory distractor. Our experiments extend the findings from Corneil et al. (2002) who assessed AV integration when visual and auditory stimuli both served as a target, and were always spatially aligned. Under such conditions the system responded according to a “*best of both worlds*” principle: as A-only saccades are typically fast but inaccurate (Fig. 4), and V-saccades are accurate but slow (Fig. 3), the AV-responses were both fast *and* accurate. These experiments demonstrated a clear integration of AV channels, whereby the interaction strength depended on the SNR of the target sound and the temporal asynchrony of the stimuli.

In the present study spatially aligned AV-targets comprised only a minority of trials (16%), while in the large majority (>80%) the auditory accessory did not provide any consistent localization cue to the system. Such a condition is arguably a more natural situation, as in typical complex environments there is no a priori knowledge about whether given acoustic and visual events originated from the same object in space.

Our data indicate that the orienting task belied a dichotomy, which was quite hard for our subjects. This was especially clear for stimuli in which the distractor preceded the visual stimulus by 75 ms (A75V condition; Figs. 5c, d and 8). In this case, the auditory input arrives substantially earlier in the CNS (by about 130 ms) and as a consequence subjects were unable to ignore the auditory distractor at short SRTs (<250 ms), as responses then appeared to be triggered by the sound. This was true for both spatially aligned (Figs. 5, 7, and 8a) and -disparate stimuli (Fig. 8e) and led to bimodal SRT distributions. A similar result for large horizontal eye-head gaze shifts was reported by Corneil and Munoz (1996) when salient AV stimuli were presented at opposite locations ( $\Delta\Phi = 180^\circ$ ,  $\Delta R = 80^\circ$ ) without an AV-background. However, the stimulus uncertainty in that study was limited, as target and distractor could occupy only two possible locations.

In line with our observations on bistability, Corneil et al. (2002) found no bimodal response distributions. Note that in their study the perceived stimulus disparity was small compared to the current study (data not shown, but mean  $\pm$  SD:  $3.3 \pm 1.4$  vs.  $19.8 \pm 15.3^\circ$ , respectively). The present study indicates that a small perceived disparity (<10°) does not elicit bistable responses (e.g., Fig. 8e, f).

Note that the height of the first SRT peak reflected the SNR of the acoustic distractor (Fig. 5c, d), which underlines our conclusion that these responses were indeed aurally guided (Fig. 8a). Interestingly, however, for the relatively rare spatially aligned condition the SRT distributions for A75V stimuli differed from the predictions of both the race model (Fig. 5e) and the bistable model (Fig. 2b) in that later responses, triggered by the visual stimulus, still had faster than visual latencies. Moreover,

even though early responses were acoustically triggered, their accuracy was better than for A-only saccades (Fig. 7). Thus, similar multisensory integration mechanisms as described by Corneil et al. (2002) also appear to operate efficiently in a rich environment that contains much more uncertainty.

Also in spatially unaligned conditions early responses were acoustically triggered and, therefore, typically ended near the location of the distractor (Fig. 8). Later responses were guided toward the visual target (Fig. 8c–f). The data from those AV stimuli thus seem to follow the predictions of the bistable model (cf. Fig. 2) much better. However, the quantitative analysis of Fig. 8c, d indicates that even in the situation of large spatial disparities the system is not driven exclusively by one stimulus modality, as responses are clearly influenced by the other modality too. Hence, a weaker form of multisensory enhancement persists that allows these responses to still outperform the unisensory-evoked saccades.

Taken together, our data show that the saccadic system rapidly accounts for the spatial–temporal relations between an auditory and visual event, and uses this information efficiently to allow multisensory integration to occur, provided the perceived spatial disparity is small. For disparities exceeding approximately 10–15°, the stimuli are treated as arising from different objects in space (Kording et al. 2007; Sato et al. 2007), which results in a bistable response mode (Fig. 8e, f). Thus, when forced to respond rapidly to a specified target, the system is prone to frequent localization errors. However, even in that case multisensory integration occurs, as the putative stimuli evoked faster and more accurate responses than their unisensory counterparts.

**Acknowledgments** We greatly acknowledge technical support of T van Dreumel and H Kleijnen. We thank R Aalbers and PM Hofman for crucial contributions to the software. We also thank prof. H Colonius for constructive comments on an earlier draft of this manuscript. Experiments were carried out in the Nijmegen Laboratory as part of the Human Frontiers Science Program (Research Grant RG-0174/1998-B; AJVO and DPM). This research was further supported by a VICI grant of the Dutch NWO/ALW (AJVO and MMVW grant nr. 805.05.003), the Canadian Institutes of Health Research (AHB and DPM), and the Radboud University Nijmegen (AJVO and MMVW).

**Open Access** This article is distributed under the terms of the Creative Commons Attribution Noncommercial License which permits any noncommercial use, distribution, and reproduction in any medium, provided the original author(s) and source are credited.

## References

- Anastasio TJ, Patton PE, Belkacem-Boussaid K (2000) Using Bayes’ rule to model multisensory enhancement in the superior colliculus. *Neural Comput* 12:1165–1187



- Bell AH, Meredith MA, Van Opstal AJ, Munoz DP (2005) Crossmodal integration in the primate superior colliculus underlying the preparation and initiation of saccadic eye movements. *J Neurophysiol* 93:3659–3673
- Binda P, Bruno A, Burr DC, Morrone MC (2007) Fusion of visual and auditory stimuli during saccades: a Bayesian explanation for perisaccadic distortions. *J Neurosci* 27:8525–8532
- Blauert J (1997) Spatial hearing: the psychophysics of human sound localization. MIT Press, Cambridge, Massachusetts
- Calvert GA, Spence C, Stein BE (2004) The handbook of multisensory processes. The MIT Press, Cambridge, MA
- Collewijn H, van der Mark F, Jansen TC (1975) Precise recording of human eye movements. *Vision Res* 15:447–450
- Colonus H (1990) Possibly dependent probability summation of reaction time. *J Math Psych* 34:253–275
- Colonus H, Diederich A (2004a) Why aren't all deep superior colliculus neurons multisensory? A Bayes' ratio analysis. *Cogn Affect Behav Neurosci* 4:344–353
- Colonus H, Diederich A (2004b) Multisensory interaction in saccadic reaction time: a time-window-of-integration model. *J Cogn Neurosci* 16:1000–1009
- Colonus H, Diederich A (2006) The race model inequality: interpreting a geometric measure of the amount of violation. *Psychol Rev* 113:148–153
- Corneil BD, Munoz DP (1996) The influence of auditory and visual distractors on human orienting gaze shifts. *J Neurosci* 16:8193–8207
- Corneil BD, Van Wanrooij M, Munoz DP, Van Opstal AJ (2002) Auditory-visual interactions subserving goal-directed saccades in a complex scene. *J Neurophysiol* 88:438–454
- Findlay JM, Walker R (1999) A model of saccade generation based on parallel processing and competitive inhibition. *Behav Brain Sci* 22:661–674 discussion 674–721
- Frens MA, Van Opstal AJ (1995) A quantitative study of auditory-evoked saccadic eye movements in two dimensions. *Exp Brain Res* 107:103–117
- Frens MA, Van Opstal AJ (1998) Visual-auditory interactions modulate saccade-related activity in monkey superior colliculus. *Brain Res Bull* 46:211–224
- Frens MA, Van Opstal AJ, Van der Willigen RF (1995) Spatial and temporal factors determine auditory-visual interactions in human saccadic eye movements. *Percept Psychophys* 57:802–816
- Gielen SC, Schmidt RA, Van den Heuvel PJ (1983) On the nature of intersensory facilitation of reaction time. *Percept Psychophys* 34:161–168
- Good MD, Gilkey RH (1996) Sound localization in noise: the effect of signal-to-noise ratio. *J Acoust Soc Am* 99:1108–1117
- Goossens HH, Van Opstal AJ (1997) Human eye-head coordination in two dimensions under different sensorimotor conditions. *Exp Brain Res* 114:542–560
- Harrington LK, Peck CK (1998) Spatial disparity affects visual-auditory interactions in human sensorimotor processing. *Exp Brain Res* 122:247–252
- Hofman PM, Van Opstal AJ (1998) Spectro-temporal factors in two-dimensional human sound localization. *J Acoust Soc Am* 103:2634–2648
- Hughes HC, Nelson MD, Aronchick DM (1998) Spatial characteristics of visual-auditory summation in human saccades. *Vision Res* 38:3955–3963
- Knudsen EI, Konishi M (1979) Mechanisms of sound localization in the barn owl (*Tyto alba*). *J Comp Physiol A* 133:13–21
- Kording KP, Beierholm U, Ma WJ, Quartz S, Tenenbaum JB, Shams L (2007) Causal inference in multisensory perception. *PLoS ONE* 2:e943
- Meijers L, Eijkman E (1977) Distributions of simple RT with single and double stimuli. *Percept Psychophys* 22:41–48
- Meredith MA, Stein BE (1986) Visual, auditory, and somatosensory convergence on cells in superior colliculus results in multisensory integration. *J Neurophysiol* 56:640–662
- Meredith MA, Nemitz JW, Stein BE (1987) Determinants of multisensory integration in superior colliculus neurons. I. Temporal factors. *J Neurosci* 7:3215–3229
- Munoz DP, Dorris MC, Pare M, Everling S (2000) On your mark, get set: brainstem circuitry underlying saccadic initiation. *Can J Physiol Pharmacol* 78:934–944
- Oldfield SR, Parker SP (1984) Acuity of sound localisation: a topography of auditory space. II. Pinna cues absent. *Perception* 13:601–617
- Ottes FP, Van Gisbergen JA, Eggermont JJ (1985) Latency dependence of colour-based target vs nontarget discrimination by the saccadic system. *Vision Res* 25:849–862
- Peck CK (1996) Visual-auditory integration in cat superior colliculus: implications for neuronal control of the orienting response. *Prog Brain Res* 112:167–177
- Press WH, Flannery BP, Teukolsky SA, Vetterling WT (1992) Numerical recipes in C: the art of scientific computing. Cambridge University Press, Cambridge, MA, USA
- Raab DH (1962) Statistical facilitation of simple reaction times. *Trans N Y Acad Sci* 24:574–590
- Robinson DA (1963) A method of measuring eye movement using a scleral search coil in a magnetic field. *IEEE Trans Biomed Eng* 10:137–145
- Sato Y, Toyozumi T, Aihara K (2007) Bayesian inference explains perception of unity and ventriloquism aftereffect: identification of common sources of audiovisual stimuli. *Neural Comput* 19:3335–3355
- Steenken R, Colonus H, Diederich A, Rach S (2008) Visual-auditory interaction in saccadic reaction time: effects of auditory masker level. *Brain Res* 1220:150–156
- Stein BE, Meredith MA (1993) The merging of the senses. MIT, Cambridge, MA
- Ulrich R, Miller J, Schrötter H (2007) Testing the race model inequality: an algorithm and computer programs. *Behav Res Meth* 39:291–302
- Zwiers MP, Van Opstal AJ, Cruysberg JR (2001) A spatial hearing deficit in early-blind humans. *J Neurosci* 21:RC142:141–145