Low-level integration of auditory and visual motion signals requires spatial co-localisation

Georg Meyer, Sophie Wuerger, Florian Röhrbein, Christoph Zetzsche

2005

Preprint:

This is an accepted article published in Exp Brain Res. The final authenticated version is available online at: https://doi.org/[DOI not available]

Exp Brain Res (2005) 166: 538–547
DOI 10.1007/s00221-005-2394-7

RESEARCH ARTICLE

Georg F. Meyer · Sophie M. Wuerger
Florian Röhrbein · Christoph Zetzsche

Low-level integration of auditory and visual motion signals requires spatial co-localisation

Received: 22 July 2004 / Accepted: 17 December 2004 / Published online: 6 September 2005

Abstract It is well known that the detection thresholds for stationary auditory and visual signals are lower if the signals are presented bimodally rather than unimodally, provided the signals coincide in time and space. Recent work on auditory–visual motion detection suggests that the facilitation seen for stationary signals is not seen for motion signals. We investigate the conditions under which *motion perception* also benefits from the integration of auditory and visual signals. We show that the integration of cross-modal local motion signals that are matched in position and speed is consistent with thresholds predicted by a neural summation model. If the signals are presented in different hemi-fields, move in different directions, or both, then behavioural thresholds are predicted by a probability-summation model. We conclude that cross-modal signals have to be co-localised and co-incident for effective motion integration. We also argue that facilitation is only seen if the signals contain all localisation cues that would be produced by physical objects.

G. F. Meyer (✉) · S. M. Wuerger
Centre for Cognitive Neuroscience, School of Psychology, University of Liverpool, Eleanor Rathbone Bldg., Bedford Street South, Liverpool, L69 7AZ, UK
E-mail: georg@liverpool.ac.uk
E-mail: sophiew@liverpool.ac.uk

F. Röhrbein · C. Zetzsche
Cognitive Neuroinformatics, School of Mathematics and Computer Science, Bremen University, Bremen, Germany
E-mail: zetzsche@informatik.uni-bremen.de

Present address: F. Röhrbein
HONDA Research Institute Europe, Care-Legien-Str. 30, 63073 Offenbach, Germany
E-mail: florian.roehrbein@honda-ri.de

Introduction

Objects or events that we perceive are usually defined by correlated information in multiple sensory modalities. Our perceptual system exploits this to facilitate detection of signals that stimulate more than one modality. Neurones in the cat superior colliculus and cortex, for instance, respond to stimulation from multiple modalities and the receptive fields of these neurones are approximately spatially aligned in all modalities (Meredith et al. 1987; Stein and Meredith 1993; Meredith and Stein 1996). The response characteristics of multi-modal neurones show non-linear interactions between the different modalities, particularly near the threshold of the respective unimodal signals. Spatio-temporally coherent stimulation leads to a response enhancement, whereas simultaneous but spatially incoherent stimulation can lead to response depression.

The neurophysiological data are mirrored by behavioural studies, which reveal enhanced perceptual processing for static visual signals in the presence of simultaneous and co-localised sounds (e.g. Frassinetti et al. 2002) and numerous cross-modal links in endogenous and exogenous spatial attention for coincident (Spence and Driver 1996, 1997) and sequential (McDonald et al. 2000) stimuli in the auditory and visual modalities.

If co-localised and co-incident signals facilitate detection by multimodal processing, then other signal features, such as common motion might be expected to have the same effect. Neurophysiological data show that, for instance, the cat superior colliculus contains neurons that respond selectively to motion signals (Wallace and Stein 1997). Functional magnetic resonance imaging (fMRI) studies have revealed several cortical areas that respond to both visual and auditory motion signals (Lewis et al. 2000) or that activation in areas traditionally considered unimodal auditory processing centres increased when subjects were presented with visual motion signals (Howard et al. 1996). The

physiological and imaging data suggest that the ''neural hardware'' to extract audio-visual motion features might reasonably be expected to be found in both peripheral and central processing areas. There is some evidence that multisensory contributions to motion perception are most pronounced when the stimuli are matched in location and time of occurrence: Soto-Faraco et al. (2002, 2004) showed that the perceived direction of apparent auditory motion between two loudspeakers is strongly modulated by simultaneous apparent visual motion. The experiments reported by Soto-Faraco et al. (2002, 2003), however, were not designed to explicitly separate effects at the perceptual level from a decision level and contradict a number of other studies that investigated audio-visual motion integration (Röhrbein and Zetzsche 2000; Meyer and Wuerger 2001; Wuerger et al. 2003; Alais and Burr 2004). The last two studies showed small reductions in the detection thresholds for bimodal motion stimuli that are consistent with the expected statistical advantage resulting from integration of two independent detection systems but failed to show the significant reduction in detection thresholds that are the hallmark of cross-modal facilitation by linear summation of the underlying unimodal signals at the perceptual level.

One possible conclusion is that the circuitry involved in the detection of static stimuli does not contribute to multi-modal motion perception (Alais and Burr 2004). This explanation implies that the motion detection subsystem does not have access to the cross-modal neural representations that have been demonstrated in static detection tasks.

An alternative explanation may be found in the spatial properties of the stimuli used in the experiments that failed to show cross-modal facilitation. Random-dot kinematograms (RDK) were used as the visual stimulus in most of these experiments. These stimuli consist of a large number of independently moving dots in a field; the proportion of dots that move in the same direction can be varied to give the illusion of global motion. Experiments that show facilitation for the detection of static stimuli typically use discrete visual signals (Spence and Driver 1997; McDonald et al. 2000; Frassinetti et al. 2002). Another important difference between the experiments that show facilitation and those that failed to do so was the nature of the auditory stimulus. In some audio-visual motion experiments the auditory signals were generated by manipulating inter-aural level differences by cross-fading a noise between two loudspeakers (Meyer and Wuerger 2001; Wuerger et al. 2003), in others the inter-aural time difference was manipulated (Alais and Burr 2004). Normal positional sound sources are defined by interaural amplitude and time differences and by position-dependent filtering of the signal due to the pinna shape (for a review see Blauert 1983). This means that the auditory signals, while providing a compelling motion illusion, did not contain all cues normally available to listeners. It might be argued that neural integration of the audio-visual signals draws directly on representations of all localisation cues and that facilitation is only likely to occur when all cues are present simultaneously. The presentation of motion signals with a full complement of auditory and visual local positional cues should therefore lead to a facilitation of motion detection. To test this hypothesis we conducted a set of motion detection experiments with real point sources for both auditory and visual signals.

We present results from two experiments. Experiment 1 shows that motion detection thresholds that are consistent with linear summation are obtained if the motion signal is generated by a series of point sources, and if the visual and auditory signal are co-localised and co-incident. If the visual and auditory signals move in different hemi-fields or in different directions, or both, then human motion detection thresholds are consistent with the probability summation seen in previous experiments. In the second experiment we measure the effective receptive field that is required for auditory and visual motion signals to be linearly summed.

Detection of audio-visual motion signals

Methods

Apparatus

The participant was seated in a chair in front of a 180° arc (Fig. 1) of 31 horizontally mounted LEDs and loudspeakers. The distance between the participant and the middle of the arc was 1.6 m. The LEDs and loudspeakers were switched on and off such that object movement was simulated. The LEDs and loudspeakers were controlled by a Tucker-Davis RP2 real-time signal processor (Tucker-Davis Technologies), which was connected to a personal computer. An additional LED, located in the centre of the arc, but slightly above the other LEDs, was on continuously and served as a fixation point.

Stimuli

The visual motion stimulus was generated by successive flashing of the LEDs. The intensity of each LED was 35 cd m^{-2}. By analogy with the visual motion signal, auditory motion was generated by applying voltage steps to the appropriate sequence of loudspeakers. The intensity of each click was 48 dB(A) at 1 cm distance from the loudspeaker. The duration of a flash or click was 2.9 ms. All motion signals moved at a speed of 30° s^{-1} and consisted of 16 unitary events. The motion signals described a 90° arc in either the left or right frontal hemi-field of the observer. The moving signals were presented in the same hemi-field (condition H+) or opposite hemi-fields (condition H−) to test whether behavioural responses reflect the response enhancement/

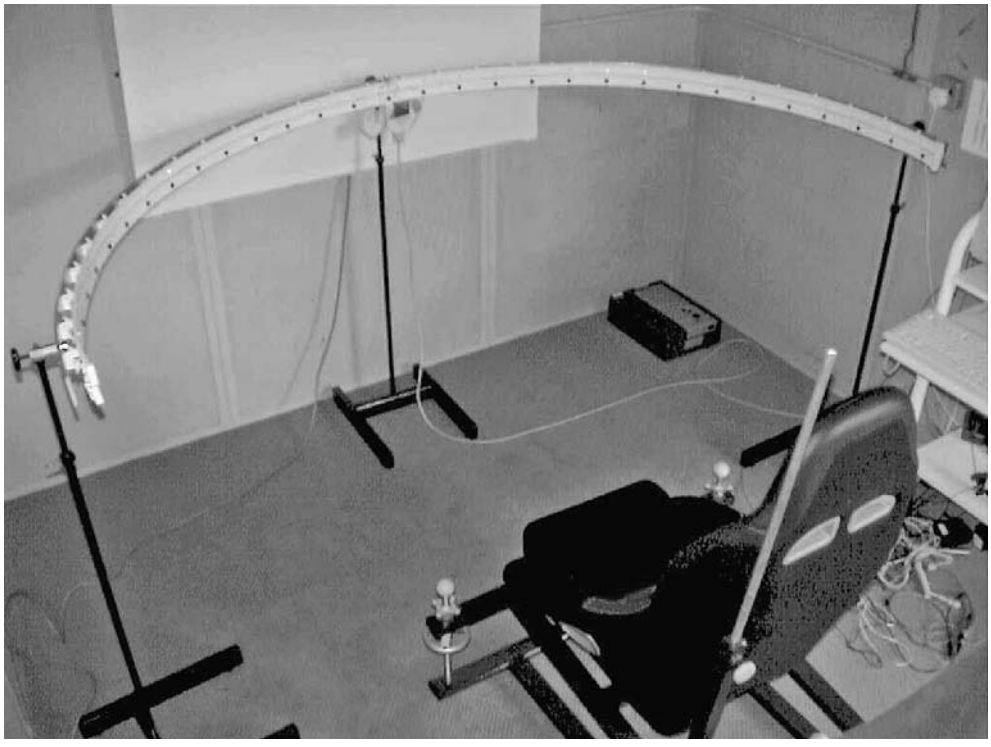

Fig. 1 Auditory–visual apparatus consisting of 31 horizontally mounted LEDs and loudspeakers. The visual motion stimulus was generated by successive flashing of the LEDs. The auditory stimulus was a click generated by a voltage step applied to the loudspeaker. In the same way as for the visual motion, auditory motion was generated by generating successive clicks. The participants were seated in a chair with a headrest in front of the auditory–visual arc and instructed to look at the fixation point

depression reported for spatially matched/mismatched stimuli (Meredith et al. 1987; Stein and Meredith 1993; Meredith and Stein 1996). The motion direction also could either be matched (D+) or mismatched (D−). A schematic diagram is given in Fig. 2; we use the "H + D−" nomenclature throughout the paper. Motion could be defined either visually, auditorily or both. The

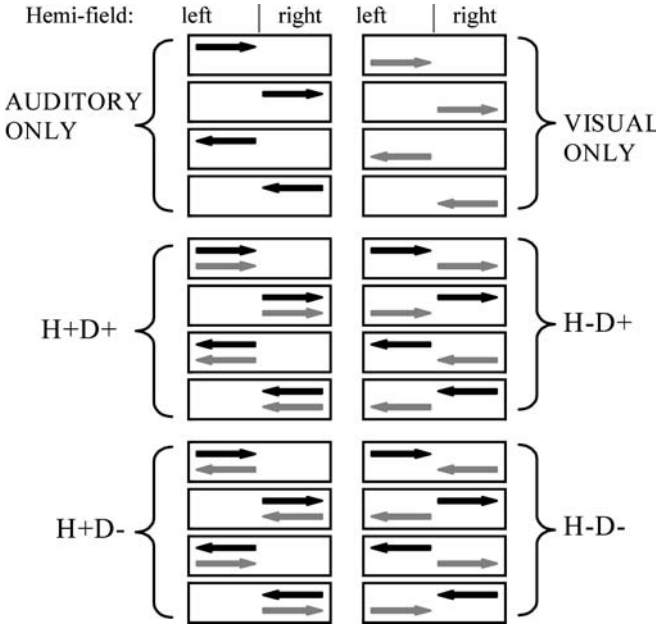

Fig. 2 Schematic diagram of the 24 experimental conditions and their labels. *H* indicates hemi-field, *D* indicates direction and +/− are used to indicate matching/mismatching motion

hemi-field, direction, and modality of the stimulus were chosen from a pseudorandom sequence. Because all component signals described a 90° arc in one hemi-field, the signals started at either the outer edges running towards the fixation point or at the fixation point and moved outwards. The motion signals were presented in a variable background of noise, generated by presenting clicks or flashes from transducers in random positions. The "noise" components were co-incident with the signal components and had the same duration and amplitude. To compute a signal-to-noise ratio (SNR) "signal" is defined as the number of clicks or flashes at successive locations while "noise" is defined as the number of randomly generated clicks or flashes. The SNR was varied in the course of the experiment to obtain the motion detection thresholds. In several precursor experiments we established that the auditory signal is affected much more by background noise than the visual signal (Section 2.2, Fig. 4). We therefore scaled the level of noise in the auditory domain to be 0.15 of the noise level in the visual trials. This ratio was kept constant for all trials in both experiments.

Procedure

All experiments were run in a darkened, sound-proof room (IAC 1402A). The participants were seated in a chair with a headrest and instructed to look at the fixation point. In a related experiment, we also recorded the eye movements with an eye-tracker (ASL 5000 Series Model 501) for two observers to ascertain that fixation can be maintained (Hofbauer et al. 2004).

In a 2IFC-task, the participant had to judge which of the two intervals contained a motion signal; one interval contained only noise (random flashes and clicks), the other contained a motion signal plus noise. The motion signal could be visual, auditory or bimodal. The participant indicated which interval contained the motion signal by pressing the appropriate button on a response box. Figure 3 shows a schematic time–space plot of visual and auditory events during one trial. The noise interval always contained the same number of clicks and flashes as the corresponding signal trial to prevent subjects from using absolute energy level differences between trials to make their decisions.

Interleaved QUEST procedures (Watson and Pelli 1983) were used to measure the unimodal (auditory and visual) and the bimodal (auditory–visual) motion detection thresholds simultaneously. The SNR in the signal trials was either modulated along a purely auditory, a purely visual, or an intermediate, auditory–visual direction (see also Fig. 4). In QUEST, at each trial a Weibull function is fitted to the relative number of correct responses as a function of the motion signal strength (SNR) and a current estimate of the threshold is obtained. The next trial is placed at the currently most likely estimate of the threshold, assuming a Gaussian prior probability density function. The final threshold estimate (after the last trial) is then the maximum likelihood estimate of the threshold based upon all data. In our experiments, threshold is defined as the SNR at which the participant can reliably (84% correct) discriminate the noise interval from the signal-plus-noise interval. Each individual threshold estimate was based on 30–40 trials. In the course of the experiment, each threshold was estimated four times for each observer.

Ten observers participated in the experiment; the total experiment consisted of four sessions. In each session six thresholds, two unimodal (auditory only, visual only) and four bimodal (H+D+, H+D−, H−D+, H−D−) thresholds were simultaneously estimated using six interleaved QUEST procedures. The starting values for all bimodal conditions were identical and based on preliminary experiments. For the bimodal conditions, the SNR for the auditory and the visual motion were updated simultaneously to maintain a fixed ratio of 0.15:1 for the auditory:visual SNR in all trials for both experiments (see "Stimuli" section). The thresholds obtained in the four sessions were averaged. Data from an initial familiarisation session were not used in the analysis.

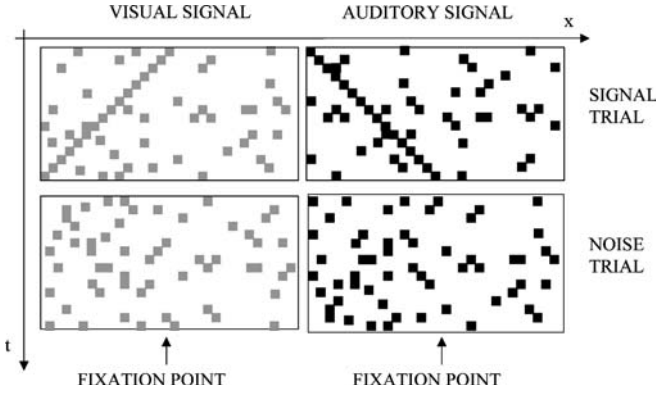

Fig. 3 Schematic diagram showing the signals as a space–time plot for the auditory and visual domain. The visual signal starts in the centre and moves to the left whereas the auditory signal runs in the same hemi-field but in the opposite direction. The noise trials were identical for the unimodal and the bimodal conditions and always consisted of auditory–visual noise. The signals were either visual alone, auditory only, or both

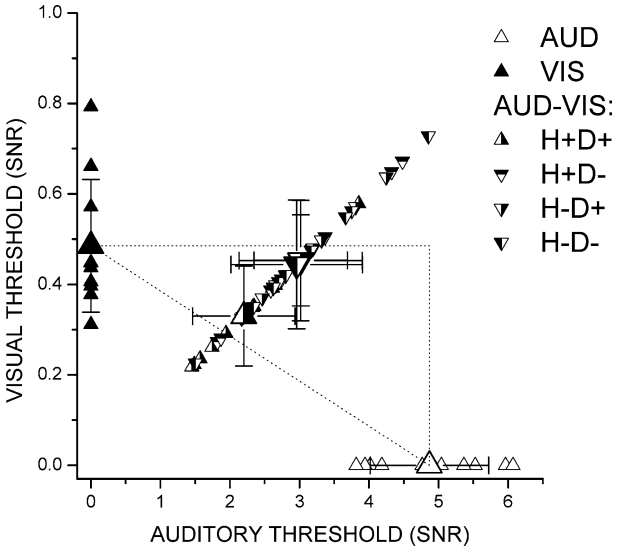

Fig. 4 Motion-detection thresholds (SNR) in the auditory–visual plane are plotted for unimodal and bimodal conditions for all observers. The auditory and the visual motion-detection thresholds in terms of SNR are plotted on the *x*-axis and *y*-axis, respectively; thresholds for the bimodal motion stimulus are shown along the intermediate direction. The visual and auditory thresholds are marked by *closed* and *open triangles*, respectively; the completely matched bimodal condition (H+D+) by a *black–white upward pointing triangle* and other non-matched bimodal conditions (H+D−, H−D+, H−D−) are indicated by different *downward pointing triangles*. The average over all observers for each AV condition is indicated by the *large symbols*. The thresholds for auditory–visual signals that are matched in direction and position of motion (H+D+) are significantly lower than the thresholds for the remaining three stimulus configurations

Results

Figure 4 shows the motion detection thresholds in the auditory–visual plane for both unimodal and bimodal conditions for all observers. The auditory and the visual motion detection thresholds in terms of SNR are plotted on the *x*-axis and *y*-axis, respectively; thresholds for the bimodal motion stimulus are shown along the intermediate direction. We did not find significant differences between the thresholds as a function of the signal location (left or right hemi-field) or for the direction of motion (leftwards or rightwards motion; compare Fig. 2 for the different conditions); the data shown in Fig. 4 are

Table 1 Unimodal auditory, unimodal visual and bimodal (H + D +, H + D−, H−D +, H−D−) motion detection thresholds (SNR) for all ten observers are shown

Observer	Unimodal		H + D +		H + D−		H−D +		H−D−	
	Aud	Vis	Aud	Vis	Aud	Vis	Aud	Vis	Aud	Vis
1	4.01	0.40	1.57	0.23	1.82	0.27	2.16	0.32	2.19	0.32
2	5.53	0.45	1.74	0.26	2.80	0.42	2.65	0.39	3.16	0.47
3	5.36	0.44	1.48	0.22	2.98	0.44	3.06	0.45	2.33	0.35
4	5.96	0.41	2.34	0.35	3.19	0.47	3.19	0.47	3.75	0.56
5	4.18	0.66	2.32	0.34	4.32	0.64	3.80	0.57	3.66	0.54
6	3.94	0.45	1.94	0.29	2.74	0.41	2.21	0.33	2.70	0.40
7	5.04	0.79	3.85	0.57	4.48	0.67	4.24	0.63	4.85	0.72
8	4.76	0.57	2.61	0.39	3.36	0.50	3.31	0.49	2.80	0.42
9	6.07	0.38	2.65	0.39	2.58	0.38	3.04	0.45	2.62	0.39
10	3.81	0.31	1.45	0.21	1.87	0.28	2.46	0.36	1.50	0.22
Mean	4.86	0.49	2.19	0.32	3.01	0.45	3.01	0.45	2.96	0.44
SD	0.85	0.15	0.73	0.11	0.88	0.13	0.67	0.10	0.94	0.14

The data show that the visual motion detection task is much more robust in background noise. In the visual domain SNR of 0.48 (or −6.74 dB) can be tolerated while in the auditory domain an SNR of 4.86 (13.73 dB) is required for reliable detection of the embedded motion stimulus.

therefore pooled across these conditions. The thresholds for unimodal stimuli differed significantly across subjects and ranged from SNRs between 3.82 and 6.07 (mean 4.86; SD 0.85) for the auditory stimuli and 0.31 and 0.79 (mean 0.49; SD 0.15) for the visual motion signals. The mean thresholds (and the standard deviations) across all observers are shown by the large symbols in Fig. 4. The data show that the visual motion detection task is much more robust in background noise. In the visual domain an SNR of 0.48 (or −6.74 dB) can be tolerated while in the auditory domain an SNR of 4.86 (13.73 dB) is required for reliable detection of the embedded motion stimulus (Table 1).

To evaluate the amount of summation between the auditory and the visual modality we compare the observers' thresholds for the bimodal stimuli with the thresholds obtained using unimodal signals. The rationale behind this summation analysis is as follows (see also Fig. 5): if both auditory and visual motion signals excite the same channel (i.e., a genuine auditory–visual motion extraction mechanism) this channel should respond much more to the bimodal stimulus than to the unimodal component signals. For a given auditory and visual motion signal the observer should therefore be more sensitive to the audio-visual stimulus than to either modality alone. This case is often referred to as "*linear summation*" or neural summation (Quick 1974; Graham 1989). If, conversely, the auditory and visual signal components excite different and independent channels (a visual channel and a separate auditory channel) then the observers' sensitivity to the bimodal stimulus should be about equal to that of the most detectable motion component. This case is usually referred to as "*independent decisions*" or "*no summation*".

Following Graham (1989) we will represent the thresholds as points in a two-dimensional summation-square plot. Figure 5 shows the predictions for four different summation rules. The auditory motion component of the bimodal stimulus (M), denoted by $M(A, AV)$ is divided by the auditory threshold, θ_a (A), and plotted on the x-axis; the visual motion component divided by the visual motion threshold is plotted on the y-axis.

$$x = \frac{M(A, AV)}{\theta_a(A)}; \quad y = \frac{M(V, AV)}{\theta_v(V)}. \tag{1}$$

In this representation, the auditory and visual motion coherence levels (SNRs) at threshold are therefore at (1,0) and (0,1), respectively. The thresholds for bimodal stimuli lie in the intermediate directions. The diagonal

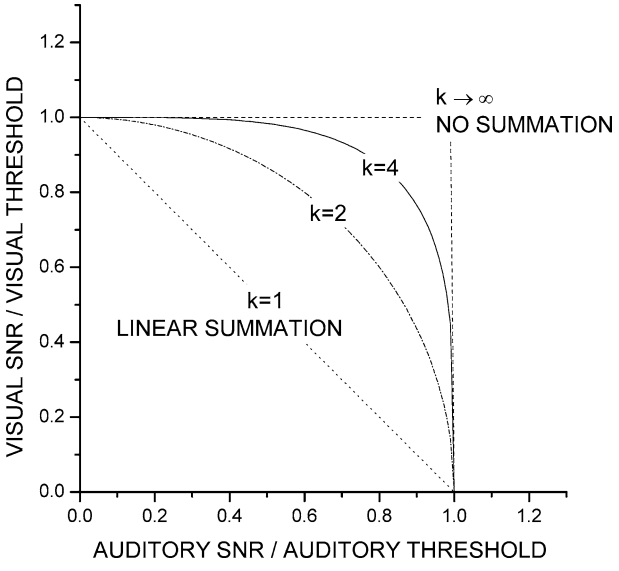

Fig. 5 Summation-square plots. Predicted motion detection contours assuming Quick's pooling model are shown. The pooling coefficients range from 1 (*linear summation*) to infinity (*no summation*). Intermediate values ranging from 3 to 5 are consistent with probability summation

line connecting the auditory (1,0) with the visual threshold (0,1) represents stimuli in which the sum of the auditory and visual motion strengths (normalised SNRs) is constant and equal to 1. An auditory–visual mechanism that linearly sums the auditory and visual motion components in a compound auditory–visual stimulus would yield thresholds that lie on this "constant-sum" line. This line is therefore labelled "*linear summation*". If the bimodal stimulus is analysed by independent auditory and visual mechanisms, then the bimodal stimulus reaches threshold whenever either the auditory or the visual motion component reaches threshold. The predicted thresholds are indicated by the dotted rectangle passing through point (1,1). This contour contains all the stimuli with either the auditory or the visual component at threshold, and is labelled "*no summation*". All points between the "linear summation" and the "no summation" contours represent bimodal stimuli at threshold with varying degrees of summation. Thresholds for the bimodal stimuli that fall outside of the rectangle indicated by "no summation" suggest a negative interaction (inhibition) between the auditory and visual channels.

To characterise the amount of summation for the auditory–visual motion signal, we use Quick's pooling formula (Quick 1974), which has been extensively used to characterise summation between visual motion mechanisms (e.g. Meese and Andersen 2002), colour mechanisms (e.g. Mullen and Sankeralli 1998) and more recently, for auditory–visual perception (Alais and Burr 2004). The Quick pooling model (Quick 1974; Graham 1989) predicts that responses of the individual analysers (i.e. the auditory and visual mechanisms) are pooled nonlinearly according to a Minkowski metric:

$$R_{\text{pool}}(AV) = (R_{\text{a}}(A)^k + R_{\text{v}}(V)^k)^{1/k} \quad (2)$$

Because the sensitivity S of a mechanism is the inverse of the threshold θ, and the response R of a mechanism is defined by the product of motion strength (M) and sensitivity (S), the x and y coordinates in Fig. 5 are therefore the responses of the auditory and visual mechanisms to a bimodal motion stimulus, respectively. A bimodal stimulus is assumed to be at threshold whenever the pooled response R_{pool} is at 1.

For a pooling coefficient k of 2, the thresholds for the compound stimulus lie on a unit circle (Fig. 5). For $k = 1$, the unimodal responses are summed linearly and the predictions lie on the contour labelled "linear summation" (Fig. 5). When k goes to infinity, the response to the bimodal stimulus is determined by the most sensitive mechanism and no summation between mechanisms occurs; in this case the thresholds for auditory–visual stimuli lie on the rectangular contour labelled "no summation". Intermediate values of k, ranging from 3 to 5, are consistent with probability summation (e.g. Tyler and Chen 2000; Meese and Andersen 2002). This means that the observer is monitoring both the auditory and the visual channel and the summation of the auditory and visual inputs is performed after a decision about the presence of motion is made in each channel. The value of the pooling coefficient k is therefore a measure for the amount of summation between the auditory and visual motion signals. Small values of k close to 1 indicate a high degree of summation; increasing values of k indicate less summation.

To evaluate the extent to which summation occurs between the auditory and the visual components in our four different conditions (matched and unmatched auditory–visual motion signals) we determined the best-fitting pooling coefficient k. Figures 6a–6d show the data and the predictions for the four different conditions (matched (a) and unmatched (b–d) auditory–visual motion information). Note that the compound thresholds for different observers will not lie on the same line through the origin, although all observers were tested with motion signals of a fixed auditory-to-visual noise ratio, because the unimodal thresholds are different for each observer (cf. Fig. 4). For the matched auditory–visual motion condition (H + D +; Fig. 6a), the best-fitting Quick pooling coefficient is close to, but always larger than 1 ($k = 1.3$) and consistent with linear summation. For the three non-matched conditions (H + D−; H−D +; H−D−; Figs. 6b–6d), the best-fitting pooling factors range from 3.8 (H−D−) to 5.1 (H + D−). These larger pooling factors are consistent with "probability summation" (Tyler and Chen 2000), which implies that the auditory and visual signals are processed independently and combined at the decision stage. We conclude that linear summation between auditory and visual motion signals only occurs if the motion signals are co-localised and co-incident (condition H + D +).

Auditory–visual receptive fields

The previous experiment shows that signals that are co-incident and co-localised are processed differently from those that are presented in different hemi-fields, or move in opposite directions, or both. For strictly co-localised motion signals human observers combine the auditory and visual inputs at an early stage (close to linear summation) for motion detection. This finding is consistent with physiological data showing that multi-sensory neurons in cat superior colliculus (Stein and Meredith 1993; Wallace and Stein 1997; Wallace et al. 1998) and in cat cortex (Wallace et al. 1992) typically have spatially matched receptive fields for auditory and visual stimuli. Visual facilitation with auditory signals was also found behaviourally by Spence and Driver (1997), who showed faster elevation discrimination for visual targets in the presence of auditory cues, and McDonald et al. (2000), who showed that a sound improves the detection of a visual signal at the same location if the delay between target and cue was less than 300 ms. Frassinetti et al. (2002) varied the spatial disparity between stationary auditory and visual signals and found significant

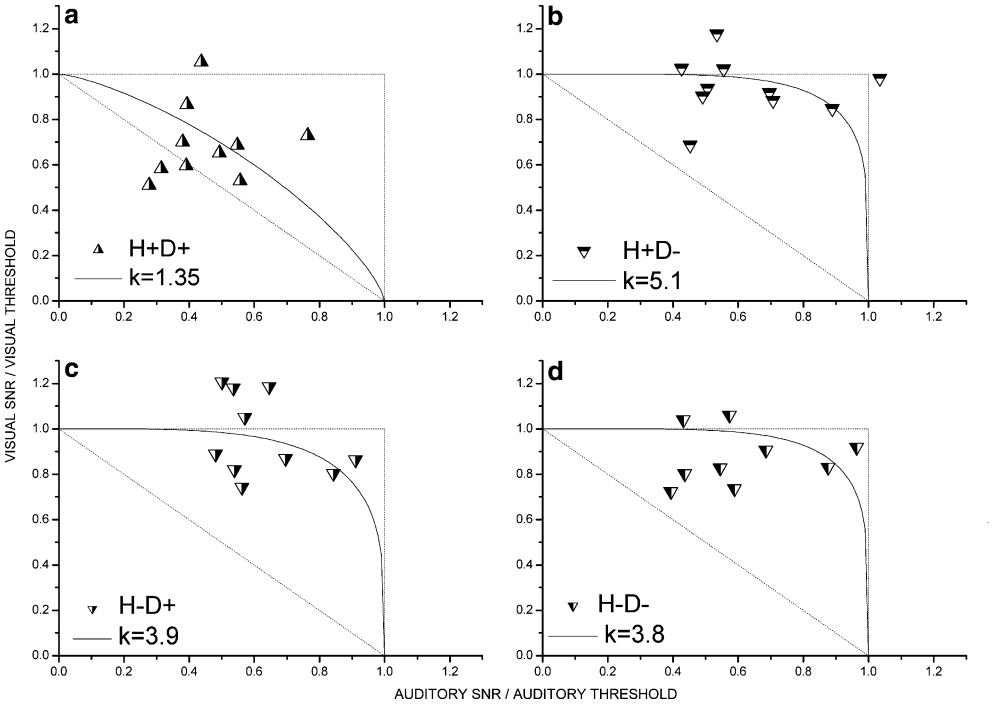

Fig. 6 Data and the predictions for the four different auditory–visual conditions (*matched* (**a**) and *unmatched* (**b–d**) auditory–visual motion information) are shown. On the *x*-axis the auditory motion SNR normalised with the auditory threshold is plotted; on the *y*-axis the visual motion SNR normalised with the visual threshold is plotted. The pooled bimodal response was fitted with the Quick pooling model

perceptual sensitivity (d') decreases for audio-visual signals that were offset by 16° visual angle (their minimum spacing between loudspeakers and LEDs). We measured this receptive field size by systematically displacing auditory and visual motion signals that moved in the same direction and at the same speed.

Methods

The equipment used, stimulus parameters and experimental conditions were the same as in the previous experiment with the following exceptions: the direction of motion was the same in all bimodal trials but auditory and visual signals were displaced in steps of 18° visual angle up to a maximum of 90°. The motion signal subtended 90°, but the start point was randomly chosen so that motion was no longer restricted to one hemi-field. A subset of five subjects took part in the experiment.

Results

Threshold data for five subjects are shown in Fig. 7. The average signal detection thresholds for the unimodal stimuli were similar to those seen in the first experiment:

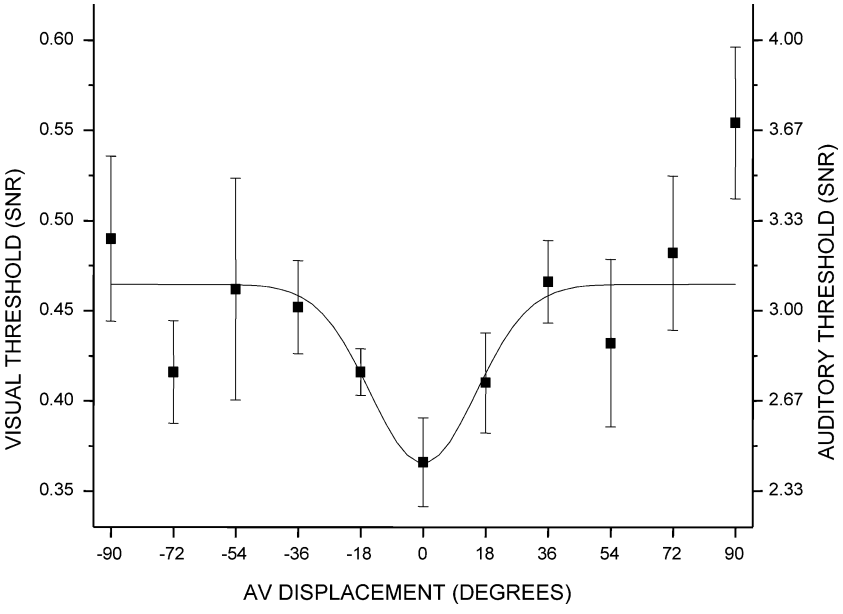

Fig. 7 Auditory–visual motion thresholds plotted as a function of the auditory–visual displacement for all five subjects. The average thresholds (in terms of SNR) are similar to those seen in the first experiment; the mean threshold SNR for the auditory unimodal signal was 5.01 ($SD = 0.285$) compared to 0.43 ($SD = 0.047$) for the visual signal. A Gaussian function was fitted to the data (*thick line*). The size of the auditory–visual receptive field (i.e. the width of the fitted Gaussian) is approximately 20° of visual angle

the mean threshold SNR for the auditory unimodal signal was 5.01 (4.86 in Exp. 1) and 0.43 (0.48 in Exp. 1) for the visual signal.

To enable systematic evaluation of the effect of displacement between the two signals, we plot the threshold SNR as a function of the displacement between visual and auditory signal. The SNR measures given on the left *y*-axis are for the visual signal whereas those on the right are the corresponding SNR estimates for the auditory signal. The squares indicate the mean threshold estimates, error bars the standard error of the means. The thick line shows a Gaussian fit (Eq. 3) through the data, weighted by the inverse standard deviation as shown in the plot. The estimated width w was 15.27 ($\pm$4.23). The other term were estimated to be $y_0 = 0.46 \pm 0.014$ and $A = -0.099 \pm 0.026$ with $R^2 = 0.64$.

$$y = y_0 + A e^{-(x^2/2w^2)} \quad (3)$$

The width w indicates that the half-height of the fitted Gaussian is located 17.97 ($\pm$ 4.98) degrees from the centre position and may be used as a crude audio-visual receptive field size measure.

Figure 8 shows the estimated pooling factors k (Eq. 2) for the five subjects as a function of the auditory–visual displacement. The estimated k value changes from 1.58 for the aligned condition to values of approximately 4 for displacements greater than 30°. The pooling factors are a measure of the amount of summation between the auditory and visual motion signals. As expected, when the auditory and visual signals are aligned, strongest summation ($k = 1.58$) is found. The data are consistent with the data found in Experiment 1 and confirm that motion detection thresholds for matched auditory and visual signals are predicted by a process close to linear summation while thresholds for mismatched stimuli are better predicted by probability summation ($k = 4.0$).

Discussion

Our experiments show that motion detection thresholds for auditory–visual signals are significantly lower for locally consistent motion signals than for signals that are not matched in location or direction, or both. The detection of well-matched auditory–visual motion signals can be explained by linear summation models while poorly matched auditory or visual motion signals yield detection thresholds that are consistent with probability summation. The experimental data complement our own previous experiments (Meyer and Wuerger 2001; Wuerger et al. 2003) and those of Alais and Burr (2004) who showed effects that are consistent with probability summation. The key difference between the current and previous experiments is the use of physical signal sources rather than auditory and visual motion illusions. An important finding, therefore, is that auditory–visual perceptual integration requires very high quality localisation cues. Alais and Burr (2004) presented a discrete moving visual target together with auditory motion signals that were generated by manipulating inter-aural time differences. In this case no facilitation of motion detection was seen. This means that for facilitation to occur it is not sufficient to have local visual signals, but that a high quality auditory spatial signal is also crucial. To our knowledge there are no data that explicitly test the integration of a global motion signal with localised auditory motion.

Our conclusions are coincident with findings by Soto-Faraco et al. (2002, 2003) who argue that their results in a cross-modal dynamic capture task, also obtained with physical signal sources at the target locations, are caused by perceptual processing rather than decision-level integration, although the experiments cannot be directly

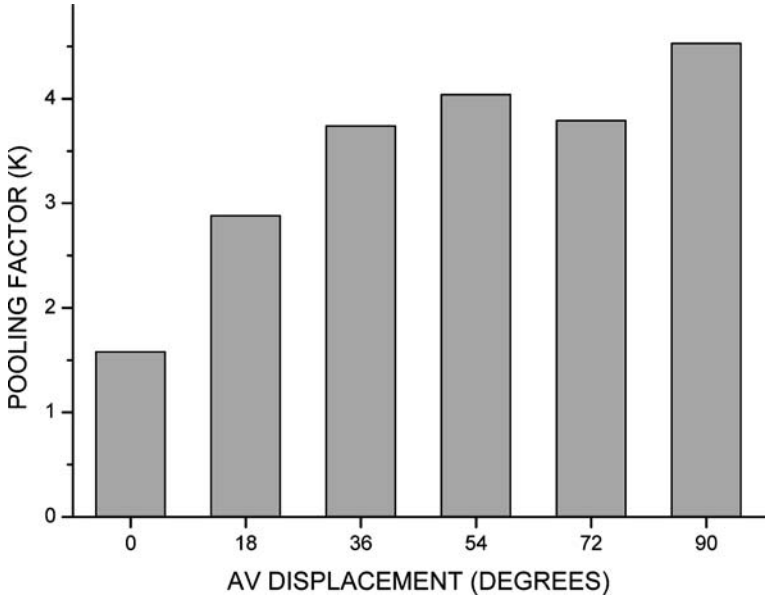

Fig. 8 The estimated pooling factors k (Eq. 2) for the five subjects as a function of the auditory–visual displacement. The estimated k-value changes from 1.58 for the aligned condition to values of approximately 4 for displacements greater than 30°. The pooling factors are a measure for the amount of summation between the auditory and visual motion signals; when the auditory and visual signals are aligned, strongest summation ($k = 1.58$) is found

compared because of the different stimulus configuration and data analysis that was used.

We show that the audio-visual stimuli have to lie within approximately 18° visual angle for effective integration to occur. This kernel size is similar to the data presented by Frassinetti et al. (2002), who show significant difference in perceptual sensitivity (d') for static audio-visual signals separated by 16° visual angle. The close match of the perceptual data for static signals and motion signals may indicate that a motion extraction subsystem has access to a mechanism that uses linear summation to integrate local auditory and visual events. The data are also consistent with physiological data that show a linear relationship between eccentricity and receptive field size with typical receptive fields of around 14° visual angle and 30° auditory angle at 0° eccentricity in adult monkeys (Wallace and Stein 2001).

This argument raises the question whether our data are based on "motion" signals or on a sequence of unrelated local events, which are easier to detect if the signal sources overlap. Our operational definition of motion is that subjects must be able to extract the instantaneous position of the signal and the speed of motion. In a separate set of experiments (Hofbauer et al. 2004) we asked subjects to "catch virtual audio-visual mice": subjects had to estimate the arrival time of an audio-visual stimulus that stopped short of a target location. This task is only possible if both the instantaneous position and motion speed of the signal are extracted. We showed that subjects are significantly more reliable at this task when bimodal stimuli rather than unimodal signals are presented. We therefore conclude that, in these experiments, it is the extraction of global motion that is facilitated when the auditory and visual signals are spatially co-localised.

Our data provide evidence for two qualitatively different integration mechanisms for audio-visual motion signals: a mechanism that is best described by a linear summation model for spatio-temporally matched motion signals and a process that is consistent with probability summation for mismatched data. Lewis et al. (2000) showed that areas activated by auditory or visual stimuli or the control of attention to auditory–visual space include the lateral prefrontal cortex, lateral parietal cortex, anterior cingulate cortex and the anterior insula. The data are not conclusive because they do not enable clear discrimination of activation that is caused by auditory or visual motion from other, task-related, activation. They do, however, show that a large number of processing sites receive simultaneous input from the visual and auditory modality. On the basis that cortical representations of auditory and visual motion signals are ubiquitous it seems at least possible that multiple representations are computed and processed by human observers. Our data also suggest there is no single hierarchical processing pathway for all audio-visual signals but what appears to be two parallel processing strategies that are active for different signal types. Sanabria et al. (2004) present data that show that the effectiveness of visual signals in the perception of auditory apparent motion is reduced if the visual component involved in the cross-modal capture is embedded in a more extensive signal. The results suggest that unimodal perceptual grouping affects cross-modal perception and, therefore, make it unlikely that the signals are processed in a strictly hierarchical fashion where bimodal signals are fused at an early stage and processed separately from unimodal signals. The results are therefore consistent with the idea that multi-sensory signals are processed by several independent modules modulated by the cues defining the signals, for example spatial position, timing, etc., and by perceptual organisation processes.

Acknowledgements This work was supported by the EU TMR projects SPHEAR and HOARSE and by the Royal Society. We are grateful to the subjects who took part in the experiments.

References

Alais D, Burr D (2004) No direction-specific bimodal facilitation for audiovisual motion detection. Cogn Brain Res 19:185–194

Blauert J (1983) Spatial hearing. MIT Press, Cambridge, MA

Frassinetti F, Bolognini N, Ladavas E (2002) Enhancement of visual perception by crossmodal visuo-auditory interaction. Exp Brain Res 147:332–343

Graham N (1989) Visual pattern analyzers. Oxford University Press, London

Hofbauer M, Wuerger SM, Meyer GF, Roehrbein M, Schill K, Zetzsche C (2004) Catching audio-visual mice: predicting the arrival time of auditory–visual motion signals. Cogn Affect Behav Neurosci 4:241–250

Howard RJ, Brammer M, Wright I, Woodruff PW, Bullmore ET (1996) A direct demonstration of functional specialization within motion-related visual and auditory cortex of the human brain. Curr Biol 6:1015–1019

Lewis JW, Beauchamp MS, DeYoe EA (2000) A comparison of visual and auditory motion processing in human cerebral cortex. Cereb Cortex 10:873–888

McDonald JJ, Teder-Sälejärvi WA, Hillyard SA (2000) Involuntary orienting to sound improves visual perception. Nature 407:906–908

Meese T, Andersen SJ (2002) Spiral mechanisms are required to account for summation of complex motion components. Vision Res 42:1073–1080

Meredith MA, Stein BE (1996) Spatial determinants of multisensory integration in cat superior colliculus. J Neurophysiol 75:1843–1857

Meredith MA, Nemitz JW, Stein BE (1987) Determinants of multisensory integration in superior colliculus neurones. I. Temporal factors. J Neurosci 10:3215–3229

Meyer G, Wuerger S (2001) Cross-modal integration of auditory and visual motion signals. Neuroreport 12:2557–2600

Mullen KT, Sankeralli MJ (1998) Evidence for the stochastic independence of the blue-yellow, red-green and luminance detection mechanisms revealed by subthreshold summation. Vision Res 39:733–745

Quick RF (1974) A vector magnitude model of contrast detection. Kybernetik 16:65–67

Röhrbein F, Zetzsche C (2000) Auditory–visual interactions and the covariance structure generated by relative movements in natural environments. In: Guidati G, Hunt H, Heiss A (eds) Proceedings of the 7th International Congress on Sound and Vibration. International Institute of Acoustics and Vibration. Kramer Technology Publishing, Munich, pp 2427–2434

Sanabria D, Soto-Faraco S, Spence C (2004) Exploring the role of visual perceptual grouping on the audiovisual integration of motion. Neuroreport 15:2745–2749

Soto-Faraco S, Lyons J, Gazzaniga M, Spence C, Kingstone A (2002) The ventriloquist in motion: illusory capture of dynamic information across sensory modalities. Cogn Brain Res 14:139–146

Soto-Faraco S, Kingstone A, Spence C (2003) Multisensory contributions to the perception of motion. Neuropsychologia 41:1847–1862

Soto-Faraco S, Spence C, Kingstone A (2004) Moving multisensory research along: motion perception across sensory modalities. Curr Direct Psychol Sci 13:29–32

Spence C, Driver J (1996) Audiovisual links in endogenous covert spatial attention. J Exp Psychol Hum Percept Perform 22(4):1005–1030

Spence C, Driver J (1997) Audiovisual links in exogenous covert spatial orienting. Percept Psychophys 59:1–22

Stein BE, Meredith MA (1993) The merging of the senses. MIT Press, Cambridge, MA

Tyler CW, Chen C-C (2000) Signal detection theory in the 2AFC paradigm: attention, channel uncertainty and probability summation. Vision Res 40:3121–3144

Wallace MT, Stein BE (1997) Development of multisensory neurons and multisensory integration in cat superior colliculus. J Neurosci 17:2429–2444

Wallace MT, Stein BE (2001) Sensory and multisensory responses in the newborn monkey superior colliculus. J Neurosci 21:8886–8894

Wallace MT, Meredith MA, Stein BE (1992) Integration of multiple sensory modalities in cat cortex. Exp Brain Res 91:484–488

Wallace MT, Meredith MA, Stein BE (1998) Multisensory integration in the superior colliculus of the alert cat. J Neurophysiol 80:1006–1010

Watson AB, Pelli D (1983) QUEST: a Bayesian adaptive psychometric method. Percept Psychophys 33:113–120

Wuerger SM, Hofbauer M, Meyer GF (2003) The integration of auditory and visual motion signals at threshold. Percept Psychophys 65:1188–1196