ELSEVIER

Brain Research Bulletin 75 (2008) 753–760

BRAIN
RESEARCH
BULLETIN

www.elsevier.com/locate/brainresbull

Research report

Tri-modal integration of visual, tactile and auditory signals for the perception of sequences of events

Jean-Pierre Bresciani, Franziska Dammeier, Marc O. Ernst*

Max-Planck-Institute for Biological Cybernetics, Spemannstrasse 38, 72076 Tuebingen, Germany

Available online 7 February 2008

Abstract

We investigated the interactions between visual, tactile and auditory sensory signals for the perception of sequences of events. Sequences of flashes, taps and beeps were presented simultaneously. For each session, subjects were instructed to count the number of events presented in one modality (Target) and to ignore the stimuli presented in the other modalities (Background). The number of events presented in the background sequence could differ from the number of events in the target sequence. For each session, we quantified the Background-evoked bias by comparing subjects' responses with and without Background (Target presented alone). Nine combinations between vision, touch and audition were tested.

In each session but two, the Background significantly biased the Target. Vision was the most susceptible to Background-evoked bias and the least efficient in biasing the other two modalities. By contrast, audition was the least susceptible to Background-evoked bias and the most efficient in biasing the other two modalities. These differences were strongly correlated to the relative reliability of each modality. In line with this, the evoked biases were larger when the Background consisted of two instead of only one modality.

These results show that for the perception of sequences of events: (1) vision, touch and audition are automatically integrated; (2) the respective contributions of the three modalities to the integrated percept differ; (3) the relative contribution of each modality depends on its relative reliability (1/variability); (4) task-irrelevant stimuli have more weight when presented in two rather than only one modality.

Keywords: Multimodal integration; Vision; Touch; Audition; Virtual environments

1. Introduction

By investigating the relationships between stimuli properties and what observers perceive, fundamental research on human perception contributes to better understanding how the brain works. But this line of research has also an important role to play in more applied domains like robotics or virtual reality. For instance, a good knowledge of the stimulus–perception relationship can allow for bypassing some technical limitations when developing virtual environments. Video monitors constitute a good illustration of this point. They cannot display real movements of objects but are nonetheless able to create an illusion of movement by successively displaying single pictures faster than human visual perception can resolve them.

Because human perception is multimodal, the investigation of the mechanisms underlying multimodal integration is of particular interest. Providing users with multimodal virtual environments enhances the immersive nature of the virtual world. This point has been exploited by the video games industry, which provides the gamers with an ever-richer stimulation of their senses (e.g., stereo vision and audition, force feedback, use of motion capture). A good understanding of the principles underlying multimodal integration is also critical in robotics. For instance, if one wants to build a robot with multiple sensors, it is essential to understand how to combine the information provided by the different sensors in the most efficient way.

When similar stimuli are presented simultaneously in two modalities, the central nervous system tends to integrate these stimuli [12,17]. Multimodal integration seems to be an automatic process because it takes place even if one modality is task-irrelevant, that is, despite the explicit instruction to focus on one modality and to ignore the other [4,6,7,9,18,24,26,29,31,32,36,42,47]. For instance, if a tone burst and a visual flash are simultaneously presented at different locations, the perceived location of the auditory stimulus is generally shifted towards the actual position of the visual stimulus [4,6]. The central nervous system co-registers the two stimuli as emanating from the same physical event,

* Corresponding author. Fax: +49 7071 601 616.
E-mail address: marc.ernst@tuebingen.mpg.de (M.O. Ernst).

0361-9230/$ – see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.brainresbull.2008.01.009

which in our example implies spatial correspondence between the two modalities. Often, such biases were observed only one-way [4,25,32,39,43]. For example, visual flashes bias the perceived position of tones, but tones barely induce any bias in the perceived location of flashes when the subjects are instructed to focus on the visual stimuli [4]. This suggests that when two modalities are automatically integrated, their relative contribution to perform the task is not the same. Which modality dominates seems to depend on the nature of the task. For spatial tasks, vision usually dominates both audition and touch [4,6,10,11,19,40,44,46], whereas the auditory modality has often been reported to dominate in the temporal domain [18,21,36,41,43,48,50].

In the present experiment, we investigated the interactions between vision, touch and audition for the perception of sequences of events. Our task was neither spatial nor temporal, but consisted in counting the number of events presented in a sequence. We tested whether vision, touch and audition are automatically combined. In each session, the subjects were instructed to focus on one modality (Target) and to ignore the other modalities (Background). For each modality, we assessed the extent to which it is biased by the other two modalities, and reciprocally, how much it biases them. We also tested whether the three modalities had the same contribution and whether a Background consisting of two modalities instead of only one would evoke larger biases. Finally, we tested whether the relative contribution of each modality depended on its relative variability. Current state-of-the-art models of multimodal integration propose that the signals provided by the different sensory channels are integrated in a statistically optimal fashion, and that the relative weight of each channel is inversely proportional to its relative variability [17]. The underlying idea is that the central nervous system takes into account the relative uncertainty of the information provided by the different sensory channels to come up with a percept that is statistically nearly optimal. Some authors proposed a plausible neural implementation of these models [13,35], notably suggesting that the distribution of neuron populations could use the firing rate variability of individual neurons to code information uncertainty in a statistically optimal way [35]. Some recent behavioral experiments suggest that weighted integration might be a generic principle applying not only to the integration of redundant sensory signals but also to the integration of task-relevant and task-irrelevant sensory channels [1,2,7,8]. We found that this is true for the perception of sequences of visual and tactile events [7] as well as for the perception of tactile and auditory events [8]. If the relative variability of a modality determines its relative contribution, we expected the least variable modality to be the least susceptible to bias and the most efficient in biasing the other modalities. By contrast, we expected the most variable modality to be the most susceptible to bias and the least efficient in biasing the other modalities.

2. Materials and methods

2.1. Subjects

Eighteen right-handed subjects (aged 20–42 years, mean = 26) participated in the experiment. None of these subjects had a history of overt sensorimotor or auditory disorder, and all had normal or corrected-to-normal vision. All subjects gave their informed consent before taking part in the experiment, which

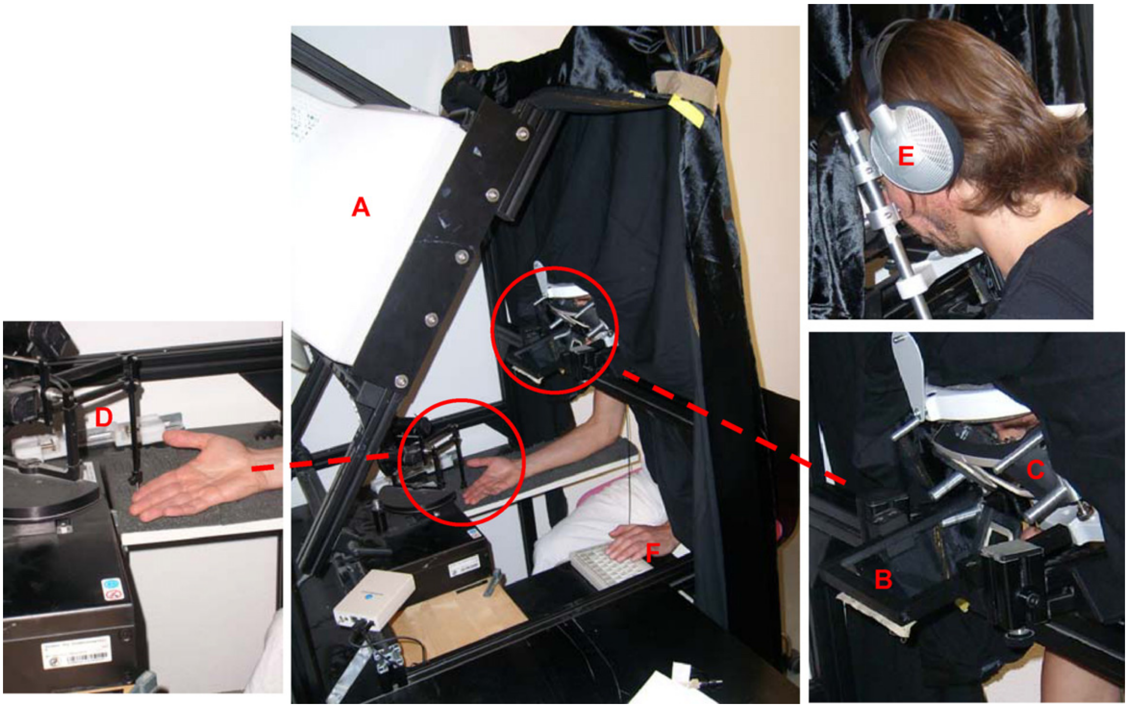

Fig. 1. Experimental set-up. The visual scene was rendered on a monitor (A) and the subject could see its reflection on a mirror (B) through stereo goggles (C). The tactile taps were delivered on the right index fingertip via a metallic pin fixed at the extremity of a PHANToM force-feedback device (D). The auditory stimuli were presented via earphones (E). The subjects gave their responses using a keypad (F).

Table 1
The nine sessions, corresponding to nine different combinations between the three modalities

Target	Background
Vision	Touch
Vision	Audition
Vision	Audition + touch
Touch	Vision
Touch	Audition
Touch	Audition + vision
Audition	Vision
Audition	Touch
Audition	Vision + touch

was performed in accordance with the ethical standards laid down in the 1964 Declaration of Helsinki.

2.2. Experimental set-up

The experimental set-up is presented in Fig. 1. The subjects were seated. Their head rested on a chin and forehead rest, whereas their right forearm and hand rested palm up at belly level on a table (72 cm high) located in front of them. A PHANToM (SensAble Technologies) force-feedback device fixed to the table was used to generate the tactile stimuli (taps of 1 N indenting subjects' skin by approximately 2 mm) via a metallic pin of 3 mm in diameter. The subjects could not see their hand or the force-feedback device. The visual scene was presented on a CRT monitor mounted up side down, and the subjects viewed its reflection in an opaque mirror (see Ref. [16] for a description of the apparatus). The visual scene consisted of a red central fixation cross (1° of visual angle) displayed for the whole duration of each session, and a white circle (1° in diameter) flashed 8.5° to the right of the central fixation cross during the trials. The visual and tactile stimuli were spatially aligned, the visual flashes being displayed at the location of the index fingertip. For the whole duration of the experiment, subjects wore earphones emitting a white-noise (71 dB) to mask any external auditory disturbance. The earphones were also used to present the auditory stimuli (beeps, 790 Hz, 74 dB). The subjects launched the trials and gave their responses with their left hand using a keypad fixed to the left of the mirror.

2.3. Procedure

The experiment was composed of nine sessions in which sequences of events were simultaneously presented in two or three of the following modalities: touch (taps), vision (flashes) and audition (beeps). For each session, the subjects were instructed to count the number of events in one modality (Target) and ignore events from the other modality(ies) (Background). Nine combinations between the three modalities were tested, i.e., each modality was the Target for three sessions, with either one, or both other two modalities as Background (see Table 1).

For each trial, a sequence of two to four events was presented in the target modality. The number of events simultaneously presented in the background modality(ies) could be: *zero* (Target presented alone), *one less* (# Background = # Target − 1), the *same number* (# Background = # Target) or *one more* (# Background = # Target + 1), for a total of 12 experimental conditions per session. When the Background consisted of two modalities, the number of events in the two background sequences was always the same. The responses were given after each trial. The subjects reported how many events they perceived in the target modality, being free to enter any number as a response. Subjects performed ten trials per experimental condition, for a total of 120 trials per session. For each session, all 12 experimental conditions were intermixed and the trials presented in a random order.

The duration of each tap, flash or beep was 50 ms, and the delay between the onsets of two successive events in the sequences was 100 ms (see Fig. 2). The delay before the onset of the target sequence was systematically adjusted so that the middle of the target and background sequences coincided in time. This adjustment allowed a maximal overlap between the target and background sequences for trials in which the amount of events in the respective sequences differed (i.e., one event less and one event more).

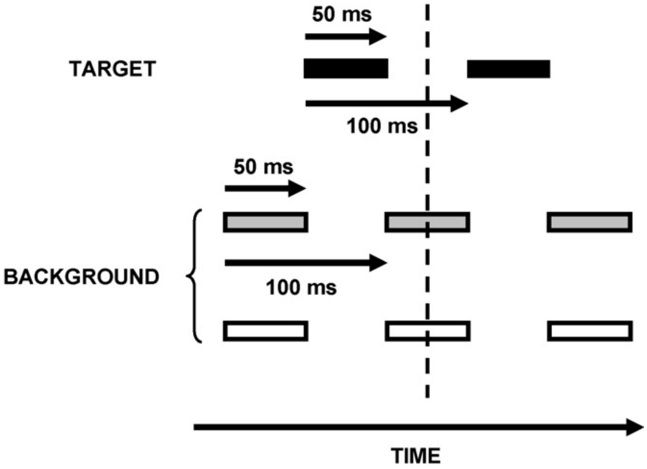

Fig. 2. Temporal profiles of the stimuli. The delay before the onset of the target sequence was systematically adjusted so that the middle of the target and background sequences coincided with respect to time. The example given here corresponds to a session in which the background consisted of two modalities and a trial in which two events were presented in the target sequence and one event more in the background sequences.

All subjects participated in the nine sessions, but in a different order for each subject (a balanced Latin square was used for the design, see Ref. [49] for details). The total experiment lasted between 120 and 150 min (about 15 min per session), which included short self-timed breaks between two successive sessions and a longer break (about 10 min) after the fifth session.

2.4. Data analysis

For testing whether the means obtained in the different experimental conditions significantly differed from one another, all statistical tests were made using repeated-measures analyses of variance (ANOVAs). When a significant effect of a main factor was observed, post hoc comparisons using Newman–Keuls tests ($p < 0.05$) were performed to determine which levels significantly differed from each other.

3. Results

3.1. Background-evoked bias for each session

For each session, we tested whether the Background biased subjects' perception. We computed Background-evoked errors for each subject. This was done by subtracting the mean of the responses obtained in the trials in which the Target was presented alone from the mean of the responses for the trials in which both Target and Background were simultaneously presented. These errors were averaged across the three Target conditions (i.e., two, three and four events presented in the target sequence) and a regression line fitted to the means. The slope of the regression line quantifies the bias induced by the Background (see Fig. 3). A slope of zero would indicate that the percept is completely determined by the Target and that the Background does not play any role (i.e., no bias), whereas a slope of one would indicate that the percept is completely determined by the Background (i.e., bias of 100%).

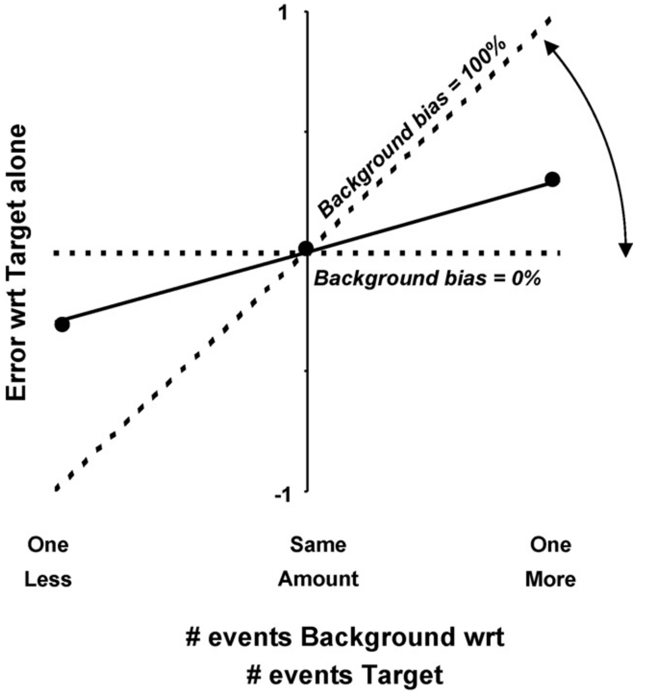

Fig. 3. For each session, Background-evoked errors with respect to the baseline percept (i.e., Target presented alone) were computed. Those were averaged across the different sequences (2, 3 and 4 events in the target sequence). The three dots represent these averaged errors for a session. For each session, a regression line was fitted to the three error values (continuous line). The slope of the regression line represents the Background-evoked bias. A slope of zero would correspond to a Background-evoked bias of zero percent, i.e., the percept depends completely on the Target. A slope of one would correspond to a Background-evoked bias of hundred percents, i.e., the percept is completely independent of the Target and completely determined by the Background.

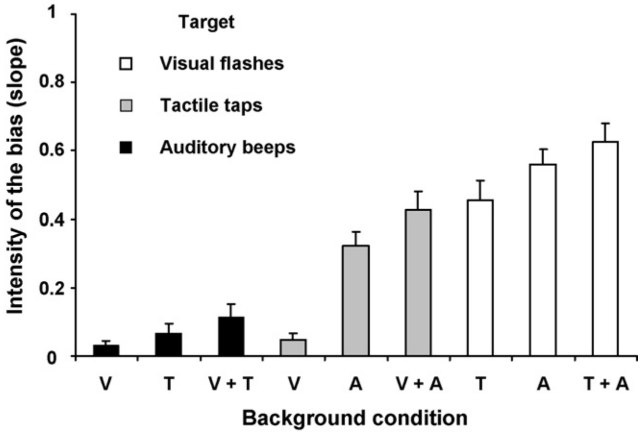

Fig. 4. Slopes of the regression lines representing the average Background-evoked bias on subjects' perception for each of the nine sessions. The letters V, T and A in legend of the *X*-axis correspond to vision, touch and audition, respectively. They indicate which modality(ies) did the Background consist of.

Fig. 4 shows the overall slope (i.e., averaged across subjects) for each of the nine sessions. The slopes are always bigger than zero, which indicates that the Background always biased subjects' responses. For each session, we tested whether this Background-evoked bias was significant. For each session, the individual response averages were entered in a 3 × 4 [number of events in the target sequence (2, 3, 4) × background condition (Target alone, one event less, same number of events, one event more)] ANOVA.

For all sessions, the perceived number of events depended on the actual number of delivered events in the target modality. Two, three and four events were always clearly discriminated (p always < 0.001). For all sessions but two, the perceived number of events also depended on the number of events presented in the background sequence(s) (see Table 2 for F and p values). The only two exceptions are the sessions in which the Background consisted of vision alone, that is, first when the Target was touch and the Background vision, and second when the Target was audition and the Background vision. Table 3 presents the detailed results of the post hoc tests.

3.2. *Susceptibility to bias of each modality*

Fig. 4 shows that the Background-evoked biases differed from one session to the other. To test whether these differences were significant, we compared the susceptibility to bias of the three modalities. The individual slopes were entered in a 3 × 3 [Target (Audition, Touch, Vision) × Background (first background modality, second background modality, combination of both background modalities)] ANOVA and we focused on the 'Target' factor.

The three modalities significantly differed from one another in terms of susceptibility to bias [$F(2, 34) = 109.42$, $p < 0.05$]. Vision (mean slope = 0.55) was significantly more susceptible to bias than touch (mean slope = 0.27), and both were significantly more susceptible to bias than audition (mean slope = 0.07).

Table 2

Detailed F and p values of the 4 × 3 [background condition (Target alone, one event less, same number of events, one event more) × number of events in the target sequence (2, 3, 4)] ANOVA testing the effect of the Background and the Target, respectively, on the perceived number of events

	Target	Vision	Vision	Vision	Touch	Touch	Touch	Audition	Audition	Audition
	Background	T	A	A + T	V	A	A + TV	V	T	V + T
Background	$F(3, 51)$	43.6	100.3	78.2	2.2	43.4	44.8	3	3.9	5.9
	p	0.000	0.000	0.000	0.095	0.000	0.000	0.052	0.013	0.002
Target	$F(2, 34)$	271.1	255.9	381	270.6	211.3	254.4	1583	469.2	760.9
	p	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

The table presents these values for each of the nine sessions.
The letters V, T and A correspond to vision, touch and audition, respectively.

Table 3
Results of the post hoc comparisons between the different background conditions (i.e., Target alone, one event less, same number of events, one event more) for each of the nine sessions

	Target Background	Vision T	Vision A	Vision A + T	Touch V	Touch A	Touch A + TV	Audition V	Audition T	Audition V + T
Background conditions	0 ↔ −1	0.012329	0.000122	0.000290	0.546085	0.001062	0.000596	0.650912	0.708091	0.678074
	0 ↔ Same	0.330397	0.785523	0.518326	0.900519	0.952708	0.841926	0.841926	0.405743	0.90051
	0 ↔ +1	0.000863	0.000135	0.000162	0.700885	0.009721	0.014350	0.490695	0.341811	0.180408
	−1 ↔ Same	0.002264	0.000131	0.000178	0.357204	0.000510	0.000845	0.693388	0.449780	0.374507
	−1 ↔ +1	0.000164	0.000164	0.000164	0.479925	0.000164	0.000164	0.387053	0.289979	0.132684
	Same ↔ +1	0.005045	0.000126	0.000241	0.865817	0.022093	0.009103	0.437516	0.568164	0.390825

The comparisons were performed using Newman–Keuls tests ($p < .05$).
The letters V, T and A correspond to vision, touch and audition, respectively.

3.3. Amplitude of the Background-evoked bias

For each modality as Target, we performed a three [Background (first background modality, second background modality, combination of both background modalities)] ANOVA to assess possible differences in the Background-evoked bias.

When the Target was vision, the Background-evoked bias was significantly stronger when the Background consisted of audition alone (slope = 0.56) and of the combination of touch and audition (slope = 0.63) than when it consisted of touch alone (slope = 0.46). When the Background consisted of touch and audition combined, the biasing effect was not significantly stronger than when it consisted of audition alone.

When the Target was touch, the Background-evoked bias was significantly stronger when the Background consisted of vision and audition combined (slope = 0.43) than when it consisted of vision alone (slope = 0.05) or audition alone (slope = 0.32). Also, the audition-evoked bias was significantly larger than the one evoked by vision.

When the Target was audition, the Background-evoked effect was stronger when the Background consisted of vision and touch combined (slope = 0.11) than when it consisted of vision alone (slope = 0.03). However, the effect induced by the combination of vision and touch was not significantly different from the one induced by touch alone (slope = 0.07).

3.4. Variability differences between the three modalities

We tested whether the three modalities were equally reliable (reliability = 1/variance) to perform the task. For each session and for each subject, we computed the standard deviation of responses for the trials in which only the target sequence was presented. For each subject, we averaged these standard deviations across the three sessions in which the Target was the same modality (after verifying in each case that the three sessions did not significantly differ from one another). This provided us with the average variability of each modality for each subject. The averaged standard deviations were entered in a 3 × 3 [modality (touch, vision, audition) × number of events in the target sequence (2, 3, 4)] ANOVA.

As shown in Fig. 5, responses variability depended on the target modality [$F(2, 34) = 48.907$, $p < .05$]. The subjects were significantly more variable in counting the visual flashes (mean standard deviation = 0.58) than in counting the tactile taps (mean standard deviation = 0.45), and in both cases significantly more variable than in counting auditory beeps (mean standard deviation = 0.25). Responses variability also depended on the number of events presented [$F(2, 34) = 10.066$, $p < 0.05$]. Subjects' responses were significantly more variable when four events were presented than when only two events were presented. There was no interaction between the modality and the number of presented events.

3.5. Correlation between the relative weight of the Background and the evoked bias

Statistical model of multimodal integration state that the relative weight of each sensory channel is proportional to its relative reliability – reliability = 1/variance ($r_i = 1/\sigma_i^2$). Under the constraint that the weights sum to 1 and that the noise of the signals is Gaussian distributed and independent, these weights can be expressed as:

$$w_i = \frac{r_i}{\sum_j r_j} \tag{1}$$

To assess whether reliability can be used to predict the biases to be expected, we tested the strength of the correlation between the relative weight of the Background and the amplitude of the evoked bias. We computed the relative weight of each modality,

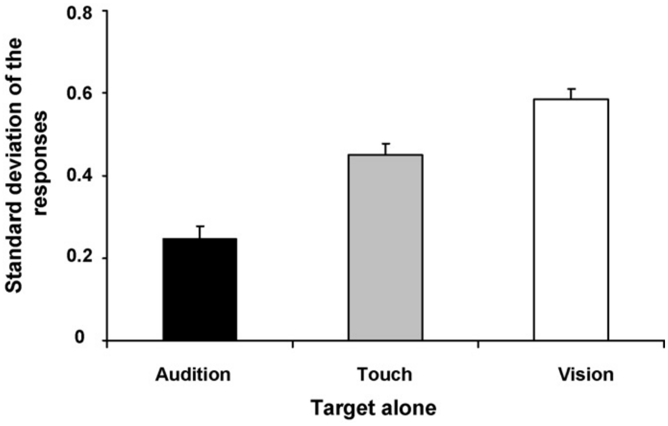

Fig. 5. Average standard deviations of the subjects' responses when the Target was presented alone (no Background).

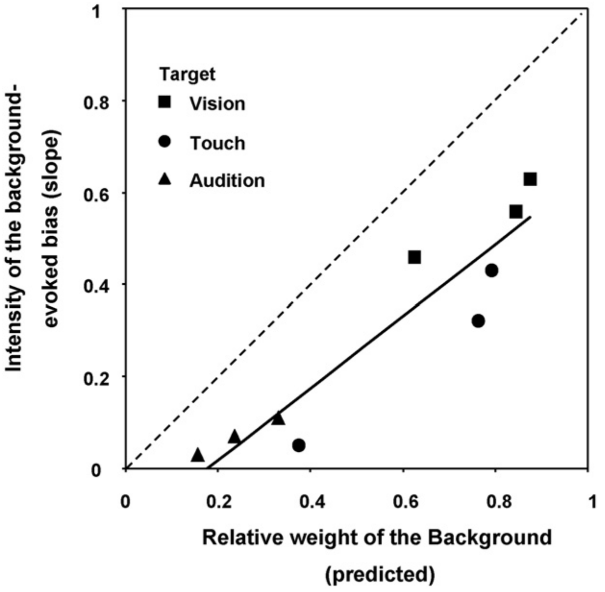

Fig. 6. Slopes representing the amplitude of the Background-evoked bias as a function of the relative weight of the Background for each of the nine sessions. The continuous line is the regression line ($y = 0.7811x + 0.1384$, $R^2 = 0.8723$) fitted to the nine data points. The dashed line ($y = x$) represents the regression line that would be obtained if the amplitude of the bias was completely determined by the relative weight of the Background.

as shown in Eq. (1), using the variance of the responses for the trials in which the Target was presented alone. In Fig. 6, the slopes representing the Background-evoked bias are plotted against the relative weight of the Background for each session. As shown by the regression line fitted to the values, the amplitude of the bias was strongly determined by the relative weight of the Background (slope of 0.78, $R = 0.93$).

4. Discussion

Vision, touch and audition were automatically integrated for the perception of sequences of events. Though the subjects were instructed to focus on one of the modalities (Target) and to ignore the other ones (Background), for all sessions but two, the perceived number of events was significantly biased by the Background. The visual modality was the most susceptible to Background-evoked bias (i.e., the most prone to be biased by the other modalities) and the least efficient in biasing the other two modalities. By contrast, the auditory modality was the least susceptible to Background-evoked bias and the most efficient in biasing the other two modalities. These results highlight a ‘hierarchy’ between the three modalities in terms of relative contribution. Interestingly, the rank of each modality in this ‘hierarchy’ was determined by its relative variance. The modality having the highest contribution was the least variable one (i.e., audition) whereas the modality having the lowest contribution was the most variable one (i.e., vision). Also, the amplitude of the Background-evoked bias was strongly correlated to the relative weight of the Background. Finally, when the Background consisted of two modalities, it induced a larger bias than when it consisted of either of the same two modalities individually. Taken together, these results suggest that the relative reliability of the sensory channels is one main factor determining their relative contribution to the integrated percept.

4.1. Automatic integration

When similar stimuli are presented in two sensory modalities simultaneously, these stimuli are merged to give rise to an integrated percept (see Refs. [12,17] for reviews). Multimodal integration seems to be automatic since it occurs even when the subjects are instructed to focus on one of the modalities and to ignore the other one [4,6,7,9,18,24,26,29,31,32,36,42,47]. One of the most famous examples of automatic integration is the ventriloquist effect [4,6]. When a tone burst and a visual flash are simultaneously presented at different locations, the perceived location of the auditory stimulus is generally shifted towards the actual position of the visual stimulus. In the present experiment, we showed for the first time an automatic integration between three modalities, namely vision, touch and audition. Our results show that these three modalities bias one another for the perception of sequences of events. This provides further evidence that the central nervous system tends to automatically integrate multimodal stimuli when these stimuli are likely to be generated by the same physical event. For perception, multimodal integration presents two advantages: it reduces the variance of the estimates [1,7,16,22] and it enhances stimulus detection [5,14,23,27,37,38]. Automatically integrating similar sensory signals has therefore a functional relevance since it allows the central nervous system to take advantage of information redundancy. At the neural level, multimodal integration probably relies on the existence of multimodal neurons whose firing frequency is more likely to increase when multiple rather than single sensory inputs are available [45]. In line with this, some cortical regions display a greater neural activation in response to multisensory stimulation than during unisensory stimulation [20,28,30,34].

4.2. Task-dependant hierarchy between the sensory modalities

Vision was the least efficient modality in biasing the other two modalities. A trend towards bias was always observed but it failed to reach significance. A direct comparison between the three modalities in terms of susceptibility to bias showed that vision was significantly more prone to bias than touch and audition. Therefore, for the perception of sequences of events, vision seems to be dominated by both touch and audition. This contrasts with the relative dominance of vision in ‘spatial tasks’ like localization [4,6,46], size estimation [19,40] or orientation estimation [10,11,44]. In our experiment, audition turned out to be the dominant modality. Audition was significantly less susceptible to bias than both touch and vision, and it induced significantly larger biases. Several studies report an auditory dominance for temporal estimates [18,21,36,41,43,48,50]. Our results show that this dominance applies to other ‘non-spatial tasks’. Touch had an intermediate contribution, dominating vision and being dominated by audition. This confirms the tendency observed in previous experiments investigating auditory-tactile [8,9,29] and visuo-tactile integration [7] in similar event-counting tasks.

4.3. Variability as a 'predictor' of the relative contribution of each modality

The dominant modality, audition, was the least variable modality to perform the task. More specifically, for the trials in which only the Target was presented (i.e., no Background), the responses were less variable when the subjects counted beeps than when they counted flashes or taps. By contrast, the most variable modality, i.e., vision, was the least efficient in biasing the other two modalities and the most susceptible to bias. The intermediate position of touch in terms of relative contribution (i.e., dominated by audition and dominating vision) corresponded to an intermediate position in terms of relative variability (i.e., more variable than audition and less than vision). In line with this pattern, we found that the Background-evoked bias was strongly determined by the relative weight of the Background. When the relative weight of the Background was low (e.g., vision alone as Background), the induced bias was weak. By contrast, strong biases were observed when the Background had a high relative weight (e.g., when the target modality was vision). Taken together, our results therefore suggest that when task-relevant and task-irrelevant sensory channels are automatically integrated, the relative contribution of each channel is determined by its relative reliability – 1/variance –, which is consistent with previous studies on the automatic integration of bimodal sensory signals [1,2,7].

4.4. Combining two modalities as Background increases the evoked bias

The induced bias was larger when the Background consisted of two modalities than when it consisted of just one of the same two modalities. For instance, when touch was the target modality, the Background-evoked bias resulting from the combination of audition and vision was significantly larger than the biases respectively evoked by audition alone and vision alone. The difference was substantial since the 'combined-bias' (slope of 0.43) was larger than the sum of the individual biases (slope of 0.37). Similarly, when audition was the target modality, the bias evoked by the combination of touch and vision was larger than the sum of the individual biases respectively evoked by touch alone and vision alone (slope of 0.11 when combined versus 0.10 when summed). The strong correlation between the weight of the background and the amplitude of the bias suggests that these increases of the bias resulted from an increased weight of the Background when it consisted of two modalities (see Fig. 6). In line with this, the increase in bias amplitude cannot be large if the increase in reliability is small. This likely explains why with audition as Target, combining touch and vision as Background failed to induce a bias that differed significantly from the bias evoked by touch alone ($p = 0.06$). In the same way, when vision was the Target, combining audition and touch did not change much the relative weight of the Background compared to audition alone. It is therefore not surprising to observe that the bimodal Background failed to significantly differ from the unimodal Background consisting of audition ($p = 0.09$).

5. Conclusion

By testing the interactions between three modalities, this experiment provided us with a richer picture of multimodal integration. Our main results can be summarized as follow: (1) vision, touch and audition are automatically integrated for the perception of sequences of events; (2) the three modalities do not equally contribute to the integrated percept; (3) the relative reliability of each modality can be used to predict its relative contribution to the percept; and (4) combining two congruent modalities that are task-irrelevant (i.e., Background) increases the Background-evoked bias compared to presenting just one task-irrelevant modality. Because vision, kinesthesia (i.e., touch and proprioception) and audition are the three modalities that are the most likely to be rendered in virtual environments, our results provide interesting insights for the design of such environments. For instance, they suggest that vision alone plays a minor role in feeling the contact with objects, at least when touch and sound are available. They also suggest that audition could be used to enhance the feeling of contact if it is appropriately coupled with touch. It could notably help increasing the perceived frequency of repeated contacts over a short period of time (e.g., when the tactile stimulation is around the saturation threshold of tactile sensors). In that respect, our results are in line with previous works highlighting the important role-played by sound in the perceived interactions with objects [3,15,24,33].

On a more general level, if one wants to exploit multimodal integration for modality substitution purposes (e.g., if the real-time rendering of a physical property poses computational difficulties) or to create illusions, our findings suggest that combining two substitution modalities would be more efficient than using only one because it would increase the reliability of the substitution modalities. Also, our results suggest that when developing multimodal virtual environments, having a protocol measuring the variability of each modality to perform a target task would allow the designers to determine which modalities are the most 'important' ones and which modalities are more negligible. For any given task, using such a protocol would provide useful guidelines for the management of the rendering resources. Finally, our results provide fundamental insights into the mechanisms of sensory fusion, which might be applicable in robotics as well. This is especially the case when multiple sensors are combined, as for instance in autonomous cars or with humanoids.

Acknowledgements

This work was supported by the Max-Planck Society and by the 5th Framework IST Program of the EU (IST-2001-38040, TOUCH-HapSys). The authors thank John Butler for helpful comments.

References

[1] D. Alais, D. Burr, The ventriloquist effect results from near-optimal bimodal integration, Curr. Biol. 14 (2004) 257–262.

[2] T.S. Andersen, K. Tiippana, M. Sams, Maximum Likelihood Integration of rapid flashes and beeps, Neurosci. Lett. 380 (2005) 155–160.

[3] F. Avanzini, P. Crosato, Integrating physically based sound models in a multimodal rendering architecture, Comput. Anim. Vir. Worlds 17 (2006) 411–419.

[4] R.I. Bermant, R.B. Welch, Effect of degree of separation of visual-auditory stimulus and eye position upon spatial interaction of vision and audition, Percept. Mot. Skills 42 (1976) 487–493.

[5] I.H. Bernstein, M.H. Clark, B.A. Edelstein, Effects of an auditory signal on visual reaction time, J. Exp. Psychol. 80 (1969) 567–569.

[6] P. Bertelson, M. Radeau, Cross-modal bias and perceptual fusion with auditory-visual spatial discordance, Percept. Psychophys. 29 (1981) 578–584.

[7] J.P. Bresciani, F. Dammeier, M.O. Ernst, Vision and touch are automatically integrated for the perception of sequences of events, J. Vis. 6 (2006) 554–564.

[8] J.-P. Bresciani, M.O. Ernst, Signal reliability modulates auditory-tactile integration for event counting, NeuroReport 18 (11) (2007) 1157–1161.

[9] J.P. Bresciani, M.O. Ernst, K. Drewing, G. Bouyer, V. Maury, A. Kheddar, Feeling what you hear: auditory signals can modulate tactile tap perception, Exp. Brain. Res. 162 (2005) 172–180.

[10] J.K. Collins, G. Singer, Interaction between sensory spatial after-effects and persistence of response following behavioral compensation, J. Exp. Psychol. 77 (1968) 301–307.

[11] R.H. Day, G. Singer, Sensory adaptation and behavioral compensation with spatially transformed vision and hearing, Psychol. Bull. 67 (1967) 307–322.

[12] B. De Gelder, P. Bertelson, Multisensory integration, perception and ecological validity, Trends Cogn. Sci. 7 (2003) 460–467.

[13] S. Deneve, P.E. Latham, A. Pouget, Efficient computation and cue integration with noisy population codes, Nat. Neurosci. 4 (2001) 826–831.

[14] A. Diederich, H. Colonius, D. Bockhorst, S. Tabeling, Visual-tactile spatial interaction in saccade generation, Exp. Brain Res. 148 (2003) 328–337.

[15] D.E. DiFranco, G.L. Beauregard, M.A. Srinivasan, The effect of auditory cues on the haptic perception of stiffness in virtual environments, Proc. ASME Dyn. Syst. Contr. Div. 61 (1997) 17–22.

[16] M.O. Ernst, M.S. Banks, Humans integrate visual and haptic information in a statistically optimal fashion, Nature 415 (2002) 429–433.

[17] M.O. Ernst, H.H. Bulthoff, Merging the senses into a robust percept, Trends Cogn. Sci. 8 (2004) 162–169.

[18] R. Fendrich, P.M. Corballis, The temporal cross-capture of audition and vision, Percept. Psychophys. 63 (2001) 719–725.

[19] S.M. Fishkin, V. Pishkin, M.L. Stahl, Factors involved in visual capture, Percept. Mot. Skills 40 (1975) 427–434.

[20] J.J. Foxe, C.E. Schroeder, The case for feedforward multisensory convergence during early cortical processing, Neuroreport 16 (2005) 419–423.

[21] J.W. Gebhard, G.H. Mowbray, On discriminating the rate of visual flicker and auditory flutter, Am. J. Psychol. 72 (1959) 521–529.

[22] S. Gepshtein, M.S. Banks, Viewing geometry determines how vision and haptics combine in size perception, Curr. Biol. 13 (2003) 483–488.

[23] S.C. Gielen, R.A. Schmidt, P.J. Van den Heuvel, On the nature of intersensory facilitation of reaction time, Percept. Psychophys. 34 (1983) 161–168.

[24] S. Guest, C. Catmur, D. Lloyd, C. Spence, Audiotactile interactions in roughness perception, Exp. Brain. Res. 146 (2002) 161–171.

[25] S. Guest, C. Spence, Tactile dominance in speeded discrimination of textures, Exp. Brain Res. 150 (2003) 201–207.

[26] S. Guest, C. Spence, What role does multisensory integration play in the visuotactile perception of texture? Int J. Psychophysiol. 50 (2003) 63–80.

[27] M. Hershenson, Reaction time as a measure of intersensory facilitation, J. Exp. Psychol. 63 (1962) 289–293.

[28] K. Hikosaka, E. Iwai, H. Saito, K. Tanaka, Polysensory properties of neurons in the anterior bank of the caudal superior temporal sulcus of the macaque monkey, J. Neurophysiol. 60 (1988) 1615–1637.

[29] K. Hotting, B. Roder, Hearing cheats touch, but less in congenitally blind than in sighted individuals, Psychol. Sci. 15 (2004) 60–64.

[30] J. Hyvarinen, A. Poranen, Function of the parietal associative area 7 as revealed from cellular discharges in alert monkeys, Brain 97 (1974) 673–692.

[31] V. Jousmaki, R. Hari, Parchment-skin illusion: sound-biased touch, Curr. Biol. 8 (1998) R190.

[32] N. Kitagawa, S. Ichihara, Hearing visual motion in depth, Nature 416 (2002) 172–174.

[33] R.L. Klatzky, D.K. Pai, E.P. Krotkov, Perception of material from contact sounds, Presence: Teleoperat. Vir. Environ. 9 (2000) 399–410.

[34] L. Leinonen, J. Hyvarinen, A.R. Sovijarvi, Functional properties of neurons in the temporo-parietal association cortex of awake monkey, Exp. Brain Res. 39 (1980) 203–215.

[35] W.J. Ma, J.M. Beck, P.E. Latham, A. Pouget, Bayesian inference with probabilistic population codes, Nat. Neurosci. 9 (2006) 1432–1438.

[36] S. Morein-Zamir, S. Soto-Faraco, A. Kingstone, Auditory capture of vision: examining temporal ventriloquism, Brain Res. Cogn. Brain Res. 17 (2003) 154–163.

[37] L.K. Morrell, Temporal characteristics of sensory interaction in choice reaction times, J. Exp. Psychol. 77 (1968) 14–18.

[38] R.S. Nickerson, Intersensory facilitation of reaction time: energy summation or preparation enhancement? Psychol. Rev. 80 (1973) 489–509.

[39] G.H. Recanzone, Auditory influences on visual temporal rate perception, J. Neurophysiol. 89 (2003) 1078–1093.

[40] I. Rock, J. Victor, Vision and touch: An experimentally created conflict between the two senses, Science 143 (1964) 594–596.

[41] R. Sekuler, A.B. Sekuler, R. Lau, Sound alters visual motion perception, Nature 385 (1997) 308.

[42] L. Shams, Y. Kamitani, S. Shimojo, Illusions. What you see is what you hear, Nature 408 (2000) 788.

[43] T. Shipley, Auditory flutter-driving of visual flicker, Science 145 (1964) 1328–1330.

[44] G. Singer, R.H. Day, Spatial adaptation and aftereffect with optically transformed vision: effects of active and passive responding and the relationship between test and exposure responses, J. Exp. Psychol. 71 (1966) 725–731.

[45] B.E. Stein, M.W. Wallace, T.R. Stanford, W. Jiang, Cortex governs multisensory integration in the midbrain, Neuroscientist 8 (2002) 306–314.

[46] R.J. van Beers, A.C. Sittig, J.J. Gon, Integration of proprioceptive and visual position-information: an experimentally supported model, J. Neurophysiol. 81 (1999) 1355–1364.

[47] A. Violentyev, S. Shimojo, L. Shams, Touch-induced visual illusion, Neuroreport 16 (2005) 1107–1110.

[48] Y. Wada, N. Kitagawa, K. Noguchi, Audio-visual integration in temporal perception, Int. J. Psychophysiol. 50 (2003) 117–124.

[49] W.A. Wagenaar, Note on the construction of diagram-balanced latin squares, Psychol. Bull. 72 (1969) 384–386.

[50] K. Watanabe, S. Shimojo, When sound affects vision: effects of auditory grouping on visual motion perception, Psychol. Sci. 12 (2001) 109–116.