Martinez-Conde, Macknik, Martinez, Alonso & Tse (Eds.)
Progress in Brain Research, Vol. 155
ISSN 0079-6123
Copyright © 2006 Elsevier B.V. All rights reserved

CHAPTER 14

Combining visual and auditory information

David Burr[1,2,*] and David Alais[3]

[1]*Dipartimento di Psicologia, Università degli Studi di Firenze, Via S. Nicolò 89, Firenze, Italy*
[2]*Istituto di Neuroscienze del CNR, Via Moruzzi 1, Pisa 56100, Italy*
[3]*Department of Physiology and Institute for Biomedical Research, School of Medical Science, University of Sydney, Sydney, NSW 2006, Australia*

Abstract: Robust perception requires that information from by our five different senses be combined at some central level to produce a single unified percept of the world. Recent theory and evidence from many laboratories suggests that the combination does not occur in a rigid, hardwired fashion, but follows flexible situation-dependent rules that allow information to be combined with maximal efficiency. In this review we discuss recent evidence from our laboratories investigating how information from auditory and visual modalities is combined. The results support the notion of Bayesian combination. We also examine temporal alignment of auditory and visual signals, and show that perceived simultaneity does not depend solely on neural latencies, but involves active processes that compensate, for example, for the physical delay introduced by the relatively slow speed of sound. Finally, we go on to show that although visual and auditory information is combined to maximize efficiency, attentional resources for the two modalities are largely independent.

Keywords: crossmodal integration; vision; audition; ventriloquist effect; flash-lag effect; attention

As Ernst and Bülthoff (2004) point out in their excellent review, the key to robust perception is the efficient combination and integration of multiple sources of sensory information. How the brain achieves this integration — both within and between sensory modalities — to form coherent perceptions of the external environment is one of the more challenging questions of sensory and cognitive neuroscience. Neurophysiologically, sensory interactions have become well documented over several decades. More recently, perceptual research combined with solid modeling is beginning to complement the neurophysiology. This chapter summarizes some recent psychophysical work on audiovisual interactions from our laboratories.

Pitting sight against sound: the ventriloquist effect

Ventriloquism is the ancient art of making one's voice appear to come from elsewhere, exploited by the Greek and Roman oracles, and possibly earlier (Connor, 2000). We regularly experience the effect when watching television and movies, where the voices seem to emanate from the actors' lips rather than from the actual sound source. The original explanations for ventriloquism (dating back to the post-Newtonian scientific efforts of early 18th century) assumed that it was based on the physical properties of sound, that performers somehow *projected* sound waves in a way to appear to emanate from their puppets, using special techniques (Connor, 2000). Only relatively recently has the alternative been considered, that ventriloquism is a sensory illusion created by our neural systems. These explanations assume that vision

---

*Corresponding author. Tel.: +39-050-3153175; Fax: +39-050-315-3210; E-mail: dave@in.cnr.it

DOI: 10.1016/S0079-6123(06)55014-9

predominates over sound, and somehow *captures* it (Pick et al., 1969; Warren et al., 1981; Mateeff et al., 1985; Caclin et al., 2002).

More recently, another approach has been suggested for combination of information. Several authors (Clarke and Yuille, 1990; Ghahramani et al., 1997; Jacobs, 1999; Ernst and Banks, 2002; Battaglia et al., 2003) have suggested and shown that multimodal information may be combined in an optimal way by summing the independent stimulus estimates from each modality according to an appropriate weighting scheme. The weights are given by the inverse of the variance (σ^2) of the underlying noise distribution (which can be assessed separately from the width of the psychometric function). For auditory and visual combination this can be expressed as

$$\hat{S} = w_{\mathrm{A}}\hat{S}_{\mathrm{A}} + w_{\mathrm{V}}\hat{S}_{\mathrm{V}} \qquad (1)$$

where $\hat{S}$ is the optimal estimate, $\hat{S}_{\mathrm{A}}$ and $\hat{S}_{\mathrm{V}}$ are the independent estimates for audition. w_{A} and w_{V} are the weights by which the unimodal estimates are scaled, and are inversely proportional to the auditory and visual variances σ_{A}^2 and σ_{V}^2

$$w_{\mathrm{A}} = 1/\sigma_{\mathrm{A}}^2, \quad w_{\mathrm{V}} = 1/\sigma_{\mathrm{V}}^2, \qquad (2)$$

Normalizing the sums of the weights to unity

$$k = 1/\sigma_{\mathrm{A}}^2 + 1/\sigma_{\mathrm{V}}^2 \qquad (3)$$

This model is “optimal” in that it combines the unimodal information to produce a multimodal stimulus estimate with the lowest possible variance (i.e., with the greatest reliability see Clarke and Yuille, 1990).

We (Alais and Burr, 2004b) tested the predictions of Eq. (1) directly by asking observers to localize in space brief light “blobs” or sound “clicks,” presented first separately (unimodally) and then together (bimodally). The purpose of the unimodal presentation was to measure the precision of these judgments under various conditions to provide estimates of variances σ_{A}^2 and σ_{V}^2. Fig. 1A shows typical results for four different stimuli: visual blobs of various degrees of blur and auditory tones. The data are fitted by cumulative Gaussian curves from which one can extract two parameters: the best estimate of perceived position $\hat{S}$ (often also referred to as the “point of subjective equality” or PSE), given by the point where the curves crosses 50%, and the threshold for making

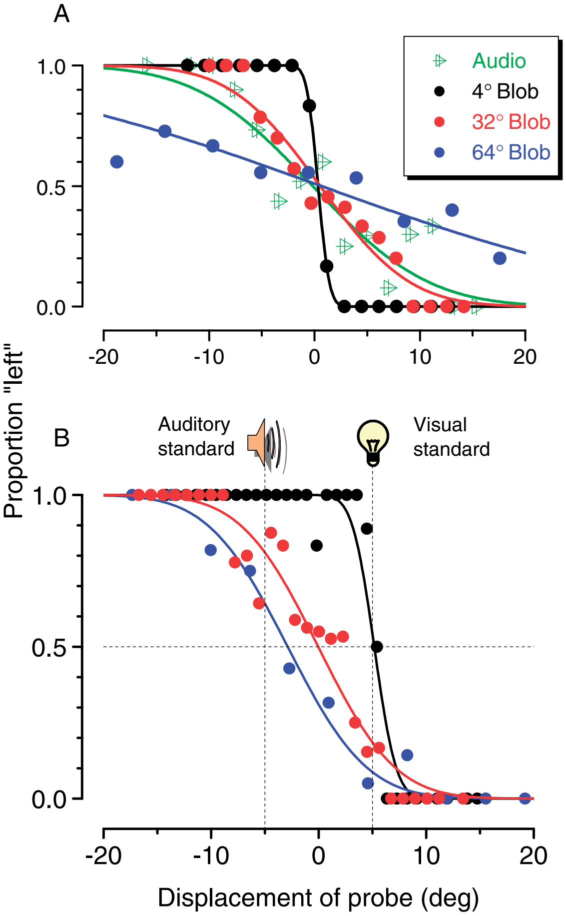

Fig. 1. (A) Unimodal psychometric functions for localization of an auditory stimulus (green), and visual Gaussian blobs of variable size. Localization for fine blobs is very good (as indicated by the steep psychometric functions), but is far poorer for very blurred blobs. Auditory localization is in between, similar to visual localization with 32° blobs. The curves are best-fitting cumulative Gaussian functions. (Reproduced with permission from Alais and Burr, 2004b.) (B) Bimodal psychometric functions for dual auditory and visual presentations. In the “conflict” presentation, the visual stimulus was displaced rightward by 5° and the auditory stimulus leftward by the same amount (as indicated by vertical lines). The 4° stimulus (black symbols) tend to follow the visual standard, the 64° stimulus (blue symbols) the auditory standard and the 32° stimulus (red symbols) falls in between. The curves are not best fits to the data, but predictions from the Bayesian model described in Eqs. (1)–(4). Modified from Alais and Burr (2004b, p. 258), Copyright, with permission from Elsevier.

the judgment, given by the width or standard deviation (σ). $\hat{S}$ was near zero for all conditions, implying that the observer, on average, saw the stimuli where it was actually displayed (at zero). However, the steepness of the curves varied considerably from condition to condition. They were steepest (small estimate of σ) for the small (4°) visual stimuli, becoming much broader for the blurred stimuli. The steepness of the auditory curves was in between, similar to the visual curve at 32°.

In the bimodal condition two different types of presentation were made on each trial, a *conflict* presentation, where the visual stimulus was displaced $+\Delta°$ and the auditory stimulus $-\Delta°$ from center, and a *probe* presentation, where the visual and auditory stimuli covaried around a mean position. Subjects were asked to judge which stimulus appeared more "rightward." Example results are shown in Fig. 1B, for $\Delta = 5°$ (meaning that the visual stimulus was displaced 5° rightward and the auditory stimulus 5° leftward, as indicated by the vertical dashed lines of Fig. 1B). The effect of the conflict clearly depended on the size of the visual blob stimuli. For 4° blobs (black symbols), the curves are clearly shifted to the right so the mean (PSE) lines up with the position of the visual stimuli. This is the classic ventriloquist effect. However, for 64° blobs (blue symbols) the reverse holds, and the curves shift leftward toward (but not quite reaching) the auditory standard. For the intermediate blur (32°, red symbols) the results are intermediate, with the PSE of the bimodal presentation falling midway between the visual and auditory standard.

The results of all the conflicts used are summarized in Fig. 2A. For each conflict and each subject, curves similar to those shown in Fig. 1B were plotted and fitted with cumulative Gaussian distributions, and the PSE (apparent coincidence of conflict and probe) was defined as the mean (50% point) of the distribution. As the example of Fig. 1B shows, for relatively unblurred visual blobs (4° blur: filled squares), vision dominated totally, while for extremely blurred blobs (128°: filled triangles), the opposite occurred, suggesting that audition dominates. At intermediate levels of blur (32°: open circles), neither stimulus dominated completely, with the points falling between the two extremes. The continuous lines are model predictions from Eq. (1), with variances σ_A^2 and σ_V^2 estimated from unimodal presentations of the auditory and visual stimuli (from curves like Fig. 1). These predictions are remarkably close to the data, providing strong evidence that Eq. (1) is applicable in these circumstances.

An even stronger test for optimal combination is that the discrimination thresholds (square root of the variances) of the bimodal presentation increases should increase

$$\sigma_{VA}^2 = \frac{\sigma_V^2 \sigma_A^2}{\sigma_A^2 + \sigma_V^2} < \min(\sigma_V^2, \sigma_A^2) \tag{4}$$

where σ_{VA} is the threshold of the combined presentation that can never be greater than either the visual or the auditory thresholds. When visual or auditory variances differ greatly, σ_{AV} will be given by the lower threshold. But when they are similar, σ_{AV} will be about $\sqrt{2}$ less than either σ_A or σ_V.

Fig. 2B shows average normalized thresholds for six observers in the crossmodal task with medium-blur levels (blob size 32°), where one expects the greatest crossmodal improvement. To reduce subject variability, all crossmodal thresholds were normalized to unity, and the visual and auditory thresholds averaged with the same normalization factor. Both visual and auditory thresholds are about 1.4 (i.e., $\sqrt{2}$) times higher than the crossmodal thresholds. The predicted averaged crossmodal thresholds (calculated by applying Eq. (4) to the individual data, and then averaging) are very close to the obtained data.

These results strongly suggest that the ventriloquist effect is a specific example of optimal combination of visual and auditory spatial cues, where each cue is weighted by an inverse estimate of its variability, rather than one modality capturing the other. As visual localization is usually far superior to auditory location, vision normally dominates, apparently "capturing" the sound source and giving rise to the classic ventriloquist effect. However, if the visual estimate is corrupted sufficiently by blurring the visual target over a large region of space, vision can become worse than audition, and optimal localization correctly predicts that sound will effectively capture sight. This is broadly consistent with other reports of integration of sensory

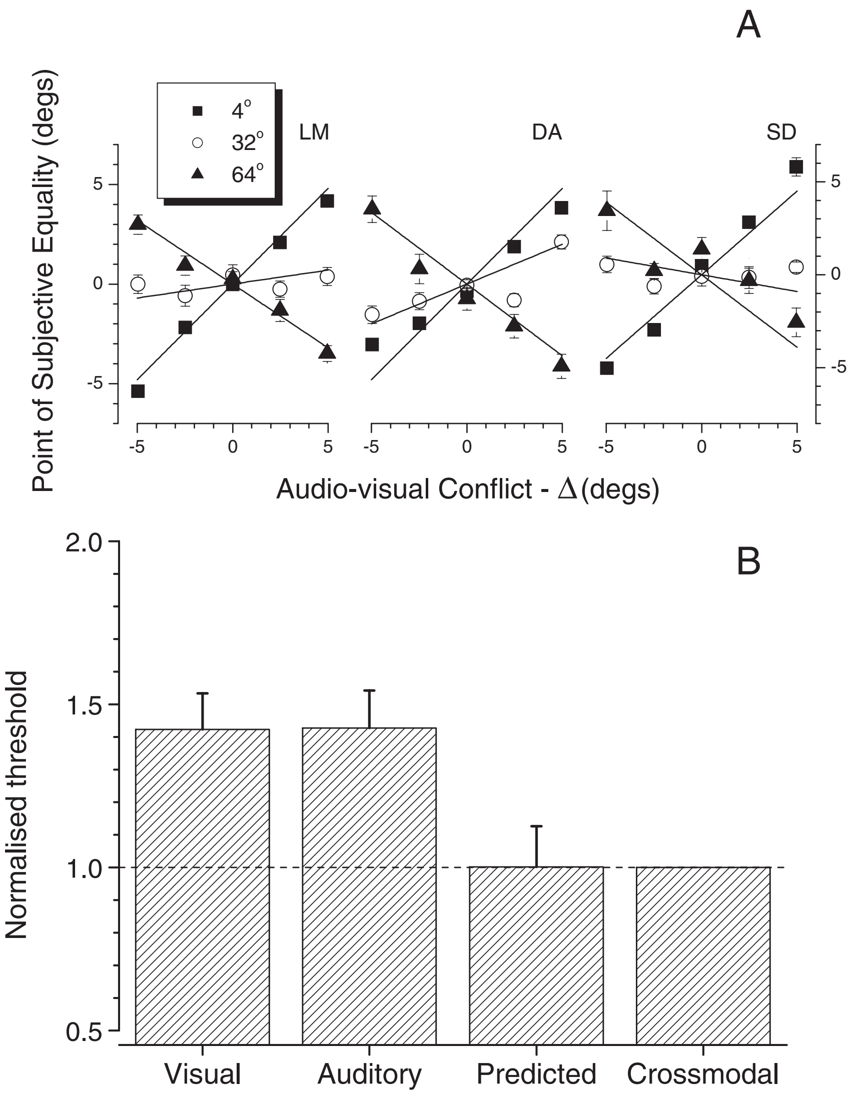

Fig. 2. (A) The effect of audiovisual conflict on spatial localization (PSE) for three different observers, and three different sizes of visual stimuli: 4° (filled squares), 32° (open circles) and 64° (filled triangles). The solid lines are the predictions of Eq. (1) using individual estimates of σ_A and σ_V for the three different sized blobs (from Fig. 1 and similar curves for the other subjects). (B) Average normalized thresholds of six subjects, for the condition where visual thresholds were similar to auditory thresholds (blob size 32°). All individual thresholds were normalized to the threshold in the crossmodal condition before averaging.
Modified from Alais and Burr (2004b, pp. 259–260), Copyright, with permission from Elsevier.

information (Clarke and Yuille, 1990; Ghahramani et al., 1997; Jacobs, 1999; Ernst and Banks, 2002; Alais and Burr, 2003). However, it differs slightly from the results of Battaglia et al. (2003) who found that vision tended to dominate more than predicted by Eq. (1): they introduced a hybrid Bayesian model to explain their effects.

Note that for auditory localization to be superior to vision, the visual targets needed to be blurred extensively, over about 60°, enough to blur most scenes beyond recognition. However, the location of the audio stimulus was defined by only one cue (interaural timing difference) and was not time varying, so auditory localization was only about one-sixth as accurate as normal hearing (Mills, 1958; Perrott and Saberi, 1990). If the effect were to generalize to natural hearing conditions, then 10° blurring would probably be sufficient. This is still a gross visual distortion, explaining why the reverse ventriloquist effect is not often noticed for spatial events. There are cases, however, when it does become relevant, not so much for blurred as for ambiguous stimuli, such as when a teacher tries to make out which child in a large class was speaking.

There is one previously reported case where sound does capture vision; this is for temporal localization where a small continuous (and peripherally viewed) light source seems to pulse when viewed together with a pulsing sound source (Shams et al., 2000; Shams et al., 2002). Furthermore, the presence of the clicks do not only make the light appear to flash, but can improve performance on visual discrimination tasks (Berger et al., 2003; Morein-Zamir et al., 2003). Although no model was offered to account for this phenomenon, it may well result from sound having far better temporal acuity than vision, resulting in the sound information being heavily weighted and appearing to capture the visual stimulus. Sounds can also modulate visual potentials in early visual areas (Shams et al., 2001), mimicking closely the modulation caused by visual stimuli, suggesting a direct interaction at an early level. Indeed preliminary evidence from our laboratories suggests that optimal, Bayesian combination of sight and sound, where the auditory temporal acuity is superior to vision, may also explain these effects (Alais and Weston, 2005; Burr et al., 2005).

An important and difficult remaining question is how the nervous system ''knows'' the variances associated with individual estimates. Must it ''learn'' these weights from experience, or could a direct estimate of variance be obtained from neural activity of a population, for example, by observing the spread of activation along a spatiotopic map? Previous studies have shown that observers can learn cue-integration strategies (Jacobs and Fine, 1999) and that the learning can be very rapid (Triesch et al., 2002). We can only guess at the neural mechanisms involved, but it is not implausible that the central nervous system encodes an estimate of measurement error along with every estimate of position, or other attribute (Ernst and Banks, 2002).

Integration of audio and visual motion

Following on from the integration of static positional cues, we asked whether auditory and visual information about motion could be effectively combined, and what are the rules of combination (Alais and Burr, 2004a). In particular, we were interested whether the combination may be ''compulsory,'' or whether observers had access to the unimodal information (see Hillis et al., 2002). Motion seemed an interesting area to study, as a key neural area involved in the multisensory combination is the superior colliculus (Stein, 1998), particular the deep layers. The superior colliculus has strong reciprocal links, via the pulvinar, with the middle-temporal (MT) cortical area (Standage and Benevento, 1983). MT is an area specialized for processing visual movement whose activity is strongly correlated with visual motion perception (Britten et al., 1992; Britten et al., 1996). MT outputs project directly to the area ventral intraparietal (VIP) where they combine with input from auditory areas to create bimodal cells with strong motion selectivity (Colby et al., 1993; Bremmer et al., 2001; Graziano, 2001). Motion perception, therefore, seemed a good area to look for strong bimodal interactions.

In order to maximize audiovisual interactions, we first measured motion detection thresholds unimodally (in two alternative forced choice) for vision and for audition, and matched them for strength. Subjects identified which interval contained the movement, without judging the direction of motion. Visual and auditory stimulus strengths were then scaled in the individual unimodal thresholds so as to be equally effective, and presented bimodally, with coherence varying together to determine the joint threshold. In separate conditions, auditory and visual stimuli moved in the same direction (and speed), or in the opposite direction (with matched speed).

Fig. 3 plots thresholds on a two-dimensional plot, with auditory coherence on the ordinate and

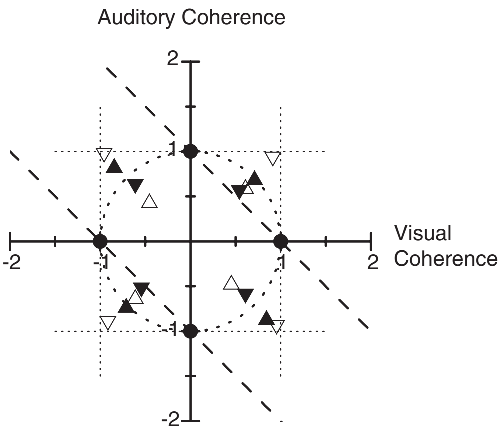

Fig. 3. Nondirectional bimodal facilitation for motion detection. The four separate subjects are indicated by different symbols on the two-dimensional plot, plotting coherence of the auditory moving stimulus against coherence of the visually moving stimulus. All thresholds are normalized so the unimodal thresholds are one. The dashed diagonal lines show the prediction for linear summation and the dashed circle for Bayesian, "statistical" summation of information. Clearly the data follow the Bayesian prediction, with no tendency whatsoever to elongate in the direction predicted by mandatory summation (–45°). Mean thresholds for same direction was 0.83, for opposite direction 0.84, with none of the observers exhibiting a significant difference.

Reproduced from Alais and Burr (2004a, p. 190), Copyright, with permission from Elsevier.

visual coherence on the abscissa. By definition, all unimodal thresholds are unity. For all observers except one (inverted triangular symbol), thresholds were lower in the bimodal than unimodal condition. However, the improvement was as good when the motion was in the opposite direction (second and fourth quadrants) as when it was in the same direction (first and third quadrants). Averaging over the four observers, mean threshold for the same-direction motion (0.83) and opposite-direction motion (0.84) were virtually identical. Clearly, the direction of the unimodal motions it was not important for bimodal motion detection.

The pattern of results is clearly not consistent with a model of linear summation of signed motion signals. The level of summation observed is too small for this (ideal prediction would be 0.5), and more importantly does not show the asymmetry toward like direction that would be expected. Perfect linear summation would follow the dashed lines oriented at –45°. Of course, this prediction is somewhat extreme, but any form of mandatory fusion should lead to an elongation of the threshold ellipse, so it is longer alone the –45° axis (where the visual and auditory directions are opposed, and should tend to annul each other). Our results give no indication whatsoever of this elongation, agreeing with Hilis et al. (2002) who demonstrated mandatory fusion within a sensory system (vision) but not between vision and touch.

The summation is, however, consistent with a statistically optimal combination of signals based on maximum likelihood estimation of Eq. (4) discussed in the previous section, and indicated in Fig. 3 by a dashed circle (Clarke and Yuille, 1990; Ghahramani et al., 1997; Jacobs, 1999; Ernst and Banks, 2002). As the auditory and visual weights were equated by equating the unimodal thresholds, the expected improvement from Eq. (4) is a factor of $1/\sqrt{2}$ ($=0.71$), not very different from the observed 0.84. Importantly, the prediction is the same for like and opposite motion, as both carry the same amount of information, although they are perceptually very distinct.

Taken together, these results show a small nondirectional gain in bimodal movement detection for bimodal motion, consistent with statistical combination, but not with a direct summation of signed audio and visual motion signals. This held true both for coherently moving visual objects and for spatially distributed motions, in central and in peripheral vision (Alais and Burr, 2004a), agreeing with two recent studies using similar methods and stimuli (Meyer and Wuerger, 2001; Wuerger et al., 2003).

Temporal synchrony — the flash-lag effect

It has long been known that the order in which perceptual events are perceived does not always reflect the order in which they were presented. For example, Titchener (1908) showed that salient, attention-grabbing stimuli are often perceived to have occurred before less salient stimuli (the "prior entry effect"). More recently, Moutoussis and Zeki (1997) showed that different attributes of the same object can appear to change at different times: if the color and direction of motion change simultaneously, color seems to lead. But perhaps the clearest example of a systematic temporal mislocalization is the so-called "flash-lag effect," first observed by MacKay (1958) and more recently revised and extensively studied by Nijhawan (1994); for review see (Krekelberg and Lappe, 2001). If a stationary disk is briefly flashed at the exact moment when a moving disk passes it, the stationary disk seems to "lag" behind the moving disk. Many explanation of the flash-lag effect have been suggested, including spatial extrapolation (Nijhawan, 1994), attention (Baldo and Klein, 1995), differential neural latencies (Purushothaman et al., 1998), spatial averaging (Krekelberg and Lappe, 2000) and "postdiction" (Eagleman and Sejnowski, 2000).

Whatever the explanation for the effect, an interesting question is whether it is specific for visual stimuli, or whether it also occurs in other senses, and crossmodally, and whether these effects could reasonably be attributed to neural latencies. We therefore measured the flash-lag effect for auditory stimuli, both for spatial motion and for spectral motion in frequency. In both cases a strong flash-lag effect was observed (Fig. 4): the stationary stimulus seemed to lag 160–180 ms behind the moving stimulus, whether the motion was in space or in frequency. This effect is in the same direction as that observed for vision, but far stronger: visual effects under the conditions of this experiments were about 20 ms. It was also possible to measure the effect crossmodally: using a visual flash as probe to a moving sound or a sound burst as probe to a moving visual stimulus. Both these conditions produced large and reliable flash-lag effects, roughly midway between the purely visual and purely auditory effects.

These results show that the flash-lag effect is not peculiar to vision, but occurs in audition, and also crossmodally. They also provide the possibility of investigating the mechanisms producing the effects, by comparing the magnitudes under the various audio and visual conditions. If the flash-lag effect were simply due to differences in neural latencies and processing time, then the relative latencies necessary to produce the results of Fig. 4 are easily calculated. As the auditory–auditory effects were the largest, the neural response to auditory motion would have to be much faster than that to an auditory flash (by about 180 ms). As the visual–visual effects were small, the response to visual motion should be only about 20 ms faster than that to a visual flash. And as the auditory–visual and visual–auditory effects were of comparable size, the visual latencies should be between the auditory motion and flash latencies. The best estimates to give the results of Fig. 4 are shown in Fig. 5A, normalizing the visual latency estimate arbitrarily to 100 ms. Fig. 5B shows recent results measuring neural delays for visual and auditory moving and stationary stimuli with three different techniques: an integration measure, perceptual alignment and reaction times (Arrighi et al., 2005). These three measures all agree quite well with each other, in suggesting that they are measuring the same thing. However, the order of the latencies measured directly is quite different from that required for the flash-lag effect. For audition, motion latencies were systematically longer than flash latencies, whereas the reverse is required for the flash-lag effect, both in audition and crossmodally.

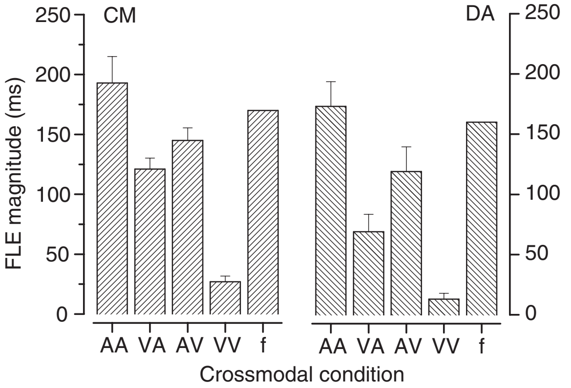

Fig. 4. Magnitude of the flash-lag effect for various auditory–visual conditions. The column indicated "*f*" refers to "motion" up and down the scales played to one ear, to which subjects had to align a tone played to the other ear. For all other bars, the first symbol refers to the modality of the moving stimulus and the second to that of the stationary "flash." Reproduced from Alais and Burr (2003, p. 60), Copyright, with permission from Elsevier.

These results reinforce previous work showing that the flash-lag effect does not result directly

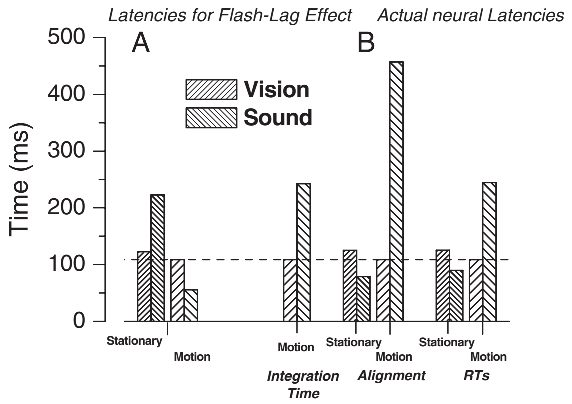

Fig. 5. (A) The relative hypothetical neural latencies necessary to account for the flash-lag data of Fig. 4, assuming simple linear accumulation of delays. Rightward hatching refers to vision, leftward to audition, sparse hatching to motion and dense hatching to stationary stimuli. Auditory motion needs to be processed the fastest, auditory "flashes" the slowest and vision in between. (B) Actual latencies measured with three different techniques: integration, perceptual alignment and reaction times (aligning all visual motion results to the reaction time data, indicated by the dashed line, so only relative latencies are shown). The results are self-consistent between the three techniques, but go in the opposite direction from those required to explain the flash-lag effect (A).

Reproduced from Arrighi et al. (2005, p. 2922), Copyright, with permission from Elsevier.

from neural latencies, but clearly reflect sensory processing strategies, possibly related to calibrating motor and sensory input (Nijhawan, 1994). It is interesting that the effects should be much larger with hearing than vision. This may be related to the fact that auditory localization of position is much less precise than visual localization (see Fig. 1). This is consistent with more recent work by Nijhawan (personal communication) showing that the flash-lag effect also occurs for touch, and is much larger when measured on the forearm (where receptive fields are large and localization imprecise) than on the finger (with small receptive fields and fine localization).

Compensating for the slow propagation speed of sound

Studies of audiovisual temporal alignment have generally found that an auditory stimulus needs to be delayed by several tens of milliseconds in order to be perceptually aligned with a visual stimulus (Hamlin, 1895; Bald et al., 1942; Bushara et al., 2001). This temporal offset is thought to reflect the slower processing times for visual stimuli. This arises because acoustic transduction between the outer and inner ears is a direct mechanical process and is extremely fast at just 1 ms or less (Corey and Hudspeth, 1979; King and Palmer, 1985), while phototransduction in the retina is a relatively slow photochemical process followed by several cascading neurochemical stages and lasts around 50 ms (Lennie, 1981; Lamb and Pugh, 1992). Thus, differential latencies between auditory and visual processing generally agree quite well with the common finding that auditory signals must lag visual signals by around 40–50 ms if they are to be perceived as temporally aligned.

Most studies of audiovisual alignment, however, are based on experiments in the near field, meaning auditory travel time is a negligible factor. Studies conducted over greater distances have produced contradictory results (Sugita and Suzuki, 2003; Kopinska and Harris, 2004; Lewald and Guski, 2004) regarding whether brain can compensate for the slow travel time of sound. We recently tested whether knowledge of the external distance of an auditory source could be used to compensate for the slow travel time of sound relative to light (Alais and Carlile, 2005). We reasoned that to compensate for auditory travel time would require a robust cue to auditory source distance, since it involves overriding the temporal difference between the signals as they arrive at the listener. We therefore used the most powerful auditory depth cue — the direct-to-reverberant energy ratio (Bronkhorst and Houtgast, 1999) — to indicate source distance.

To create a suitable sound stimulus, we recorded the impulse response function of a large concert auditorium (the Sydney Opera House) and convolved it with white noise. This stimulus sounded like a burst of white noise played in a large reverberant environment (Fig. 6). It began with a direct (i.e., anechoic) portion lasting 13 ms, followed by a long reverberant tail that dissipated over 1350 ms. To vary the apparent distance of the sound burst, we varied the amplitude of the initial part of the

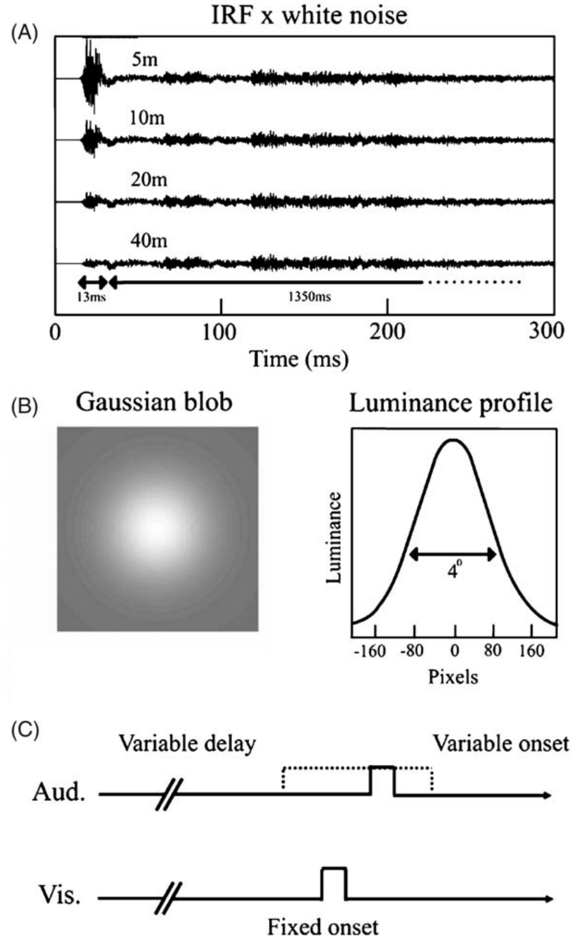

Fig. 6. The stimuli and procedures used to measure visual-acoustic synchrony. (A) The impulse response function on the top row (5 m) is the original function recorded in the Sydney Opera House convolved with white noise. The direct sound is the initial portion of high amplitude, and the long tail reverberant signal, which lasted 1.35 ms (identical for all four stimuli). Because the direct-to-reverberant energy ratio is a very strong cue to auditory source distance, attenuating the direct portion by 6 dB (halving amplitude) simulates a source distance of 10 m (see *Methods*). Further 6 dB attenuations simulated auditory distances of 20 and 40 m. (B) The visual stimulus was similar to that shown (*Left*), a circular luminance patch that was presented for 13 ms. The spatial profile of the stimulus (*Right*) was Gaussian with a full half-width of 4° of visual angle. (C) The onset of the auditory stimulus (*Upper*) was varied by an adaptive procedure to find the point of subjective alignment with the visual stimulus (*Lower*). A variable random period preceded the stimuli after the subject initiated each trial.

Reproduced from Alais and Carlile (2005, p. 2245), Copyright, with permission.

stimulus, while leaving the reverberant tail fixed for all simulated depths. Since the energy ratio of the early direct portion to the later reverberant tail is a powerful cue to auditory depth, we could effectively simulate a situation in which a sound source was heard at various distances in a constant reverberant environment, in a darkened high-fidelity anechoic chamber. To measure perceived audiovisual alignment, a brief spot of was light flashed on a dark computer screen and served as a temporal reference point. The sound onset was advanced or retarded in time using an adaptive staircase method until the onset of the sound burst was perceived to be synchronous with the light flash.

The original recording in the auditorium was made 5 m from the sound source, and successive 6 dB scaling of the early direct portion simulated stimuli at 10, 20 and 40 m (see Fig. 6a). In enclosed reverberant environments, the direct-to-reverberant energy ratio is the strongest cue to auditory sound source distance because the incident level decreases by 6 dB with each doubling of distance while the level of the reverberant tail is approximately invariant (Bronkhorst and Houtgast, 1999; Zahorik, 2002; Kopinska and Harris, 2004).

The results were clear: the point of subjective alignment of auditory and visual stimuli depended on the source distance simulated in the auditory stimulus. Sound onset times had to be increasingly delayed to produce alignment with the visual stimulus as perceived acoustic distance increased (Fig. 7A). Best-fitting linear functions describe the data well, with slopes varying between observers from 2.5 to 4.2 ms/m, with the average (3.2 ms/m, shown by the dotted line of Fig. 7B) approximately consistent with the delay needed to compensate for the speed of sound (2.9 ms/m at 20°C, indicated by the dashed line). These results suggest that subjects were attempting to compensate for the travel time from the simulated source distance using a subjective estimate of the speed of sound.

Various controls were performed to show that the reverberant tail of the sound wave was essential for the subjective audiovisual alignment to shift in time (Alais and Carlile, 2005). In a further control, the observers' attention was focused on the onset burst by requiring them to make speedier responses (slow responses were rejected). Under this condition

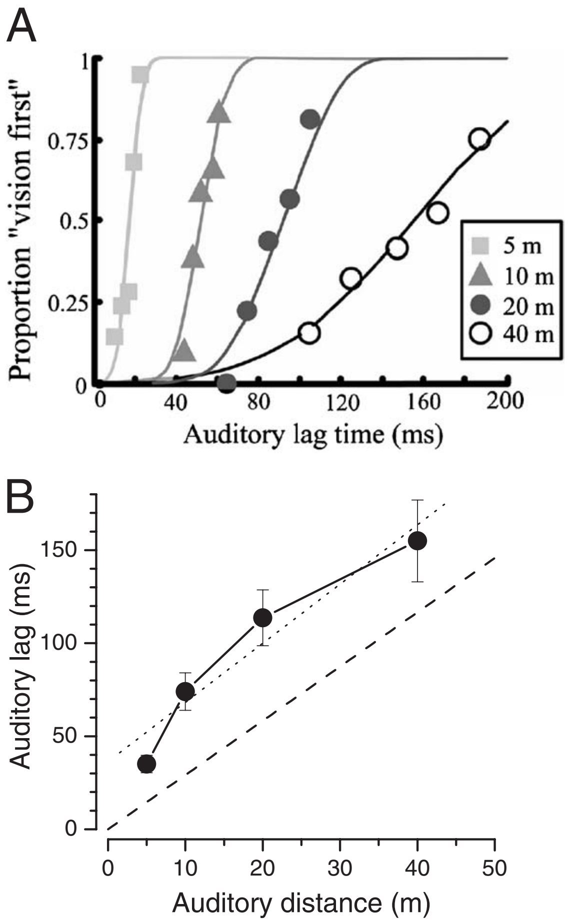

Fig. 7. (A) Psychometric functions for one observer at each of the four simulated auditory distances plotting the proportion of trials in which the visual stimulus was judged to have occurred before the auditory stimulus, as a function of the delay of the auditory stimulus. From left to right, the curves represent the 5-, 10-, 20- and 40-m conditions. The abscissa shows time measured from the onset of the visual stimulus. (B) Average points of subjective audiovisual alignment (the half-height of the psychometric functions) for four observers at each of the four auditory distances. As auditory distance simulated by the direct-to-reverberant energy ratio increased, the auditory stimulus was perceptually aligned with earlier visual events, consistent with subjects using the energy ratio in their alignment judgments. The dotted line shows the best-fitting linear regression to the data. The slope of the regression is 3.2 ms/m, consistent with the speed of sound (2.9 ms/m at 20°C, indicated by the lower dashed line).
Reproduced from Alais and Carlile (2005, p. 2245), Copyright, with permission.

(where the reverberant tail is not attended) there is no systematic variation across auditory depth, showing that use of this cue is strategic and task dependent, rather than an automatic integration.

The essential finding from these experiments is that the brain is able to compensate for the fact that, with increasing source distance, the acoustic signal arising from a real bimodal event will arrive at the perceiver's head at progressively later times than the corresponding visual signal. These studies clearly refute any simple account of audiovisual alignment based solely on neural latencies, which would predict a common auditory lag for all simulated source distances, determined by the differential neural processing latencies for vision and audition. However, we show that the point of subjective alignment became systematically delayed as simulated auditory distance increased. Thus, the data suggest an active, interpretative process capable of exploiting auditory depth cues to temporally align auditory and visual signals at the moment they occur at their external sources.

This process could be termed "external" alignment, in contrast to "internal" alignment based on time of arrival and internal latencies. Because external alignment requires the brain to ignore a considerable temporal asynchrony between two neural signals (specifically, the late arrival of the auditory signal), it is unlikely to do so unless there is a robust depth cue to guide it. The direct-to-reverberant energy ratio appears to be a powerful enough cue to permit this, provided it is relevant to do so. Without a reliable depth cue, the brain seems to default to aligning signals internally, demonstrating flexibility in determining audiovisual alignment. External alignment would require knowledge of source distance and speed of sound. The direct-to-reverberant energy ratio provides a reliable auditory distance cue, and listeners presumably derive an experience-based estimate of the speed of sound, which is validated and refined through interaction with the environment.

Crossmodal attention

With the environment providing much competing input to the sensory system, selecting relevant information for further processing by limited neural resources is important. Cells in the deep layers of the superior colliculus play an important role in exogenous attention. However, attention can also be deployed voluntarily (endogenous attention) to select certain stimuli from the array of input stimuli (Desimone and Duncan, 1995). Attentional selection improves performance on many tasks, as limited cognitive resources are allocated to the selected location or object to enhance its neural representation. This is true both for tasks that may be considered to be "high-level" and for those considered "low level" (for review see Pashler, 1998).

Evidence from neurophysiology, neuropsychology and neuroimaging suggests that attention acts at many cortical levels, including primary cortices. Neuroimaging and single-unit electrophysiology point to attentional modulation of both V1 and A1 (Woodruff et al., 1996; Grady et al., 1997; Luck et al., 1997; Brefczynski and DeYoe, 1999; Gandhi et al., 1999; Jancke et al., 1999; Kanwisher and Wojciulik, 2000; see also Corbetta and Shulman, 2002).

Some psychophysical studies also show cross-modal attentional effects. For example, shadowing a voice in one location while ignoring one in another is slightly improved by watching a video of moving lips in the shadowed location (Driver and Spence, 1994), and performance can be worsened by viewing a video of the distractor stream (Spence et al., 2000). Also, precuing observers to the location of an auditory stimulus can also increase response speed to a visual target, and vice versa (Driver and Spence, 2004). On the other hand, several studies from the older psychological and human factors literature show substantial independence between visual and auditory attention (Triesman and Davies, 1973; Wickens, 1980), and some more recent studies also point in this direction (Bonnel and Hafter, 1998; Ferlazzo et al., 2002). In addition, the "attentional blink" (the momentary reduction in attention following a perceptual decision) is modality specific, with very littler transfer between vision and audition (Duncan et al., 1997).

Overall, the evidence relating to whether attention is supramodal or whether it exists as a separate resource for each modality is equivocal. We therefore measured basic discrimination thresholds for low-level auditory and visual stimuli while dividing attention between concurrent tasks of the same or different modality. If attention is a single supramodal system, then a secondary distractor task should reduce performance equally for intramodal and extramodal distractor tasks. However, if there are separate attentional resources for vision and audition, then extramodal distractors should not impair performance on the primary task. Our results suggest that vision and audition have their own attentional resources.

We measured discrimination thresholds for visual contrast and pitch, initially on their own, then while subjects did a concurrent secondary task that was either intramodal or extramodal. The secondary (distractor) task for the visual modality was to detect whether one element in a brief central array of dots was brighter than the others, and the secondary task in audition was to detect whether a brief triad of tones formed a major or a minor chord. Stimuli for the secondary tasks had a fixed level of difficulty (1 standard deviation above threshold level, as determined in a pilot experiment).

Fig. 8 shows psychometric functions from one observer showing performance on the primary visual task (contrast discrimination, left-hand panel) and on the primary auditory task (frequency discrimination, right-hand panel). In each panel, filled circles represent performance on the primary task when measured alone, while the two other curves show performance on the primary task when measured in the dual task context. The filled squares in each panel show primary task performance measured in the presence of a concurrent intramodal distractor task. The psychometric functions in this case are shifted to the right, showing a marked increase in the contrast (or frequency) increment required to perform the primary task. For all subjects, increment thresholds were at least twofold larger for intramodal distractors, and as much as fivefold. The critical condition is shown by the open triangles. These show primary task performance measured when the distractor task was extramodal. Psychometric functions in this case are very similar to those obtained without any distractor task (filled circles) indicating that for both audition and vision, primary

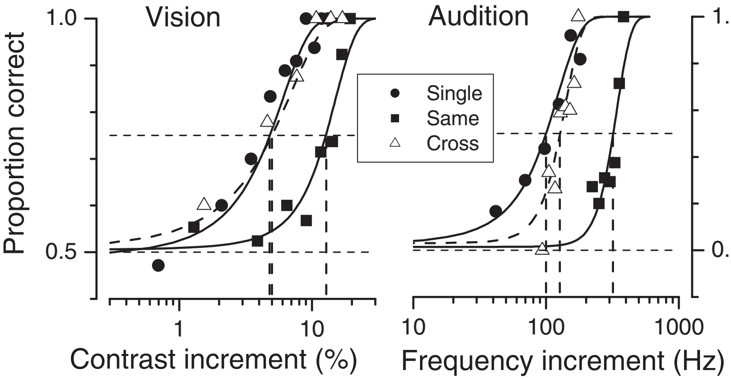

Fig. 8. Examples of psychometric functions for visual contrast and auditory frequency discriminations for one naïve observer (R.A.). The filled circles show the thresholds for the primary task alone, the filled squares when performed together with the secondary task in the same modality and the open triangles when performed with the secondary task in the other modality. Chance performance was 50% (lower dashed line). The curves are best-fitting cumulative Gaussians, from which thresholds were calculated (taken as the 75% correct point, indicated by the vertical dashed lines). The secondary task in the same modality clearly impeded performance, shifting the psychometric functions toward higher contrasts and frequencies, without greatly affecting their slope or general form. In this experiment the secondary tasks were adjusted in difficulty to produce 92% correct performance when presented alone ($d' = 2$).

task performance was largely unaffected by a competing task presented to another modality. Importantly, the psychometric functions remained orderly during the dual tasks, without decreasing slope or increasing in noise, implying a real change in the threshold limit. A marked change in slope or noisiness would have suggested that the subjects were "multiplexing" and attempting to alternate between tasks from trial to trial. This would have compromised their performance on the primary task and produced noisier data with a shallower slope.

Fig. 9 summarizes the primary thresholds in the dual-task conditions for three observers. The dual-task thresholds are shown as multiples of the primary thresholds that were measured in the single-task conditions (i.e., filled circles of Fig. 8), so that a value of 1 (dashed line) would indicate no change at all. In all cases secondary tasks that were intramodal raised primary thresholds considerably, while the extramodal secondary tasks had virtually no effect. The average increase in primary threshold produced by intramodal distractors was a factor of 2.6 for vision and a factor of 4.2 for audition, while the average threshold increase produced by extramodal distractors was just 1.1 for vision and 1.2 for audition.

The final cluster of columns in Fig. 9 shows the same data averaged over observers. The large effects of intramodal distractors are clear. Statistical tests on the two extramodal conditions (the two middle columns) showed that the mean increase in the primary auditory threshold produced by the extramodal (visual) distractor was statistically significant ($p = 0.002$); however, the mean increase in the primary visual threshold produced by the extramodal (auditory) distractor was not significantly greater than 1.0 ($p > 0.05$).

The results of these experiments clearly show that basic auditory and visual discriminations of the kind used here are not limited by a common central resource. A concurrent auditory task dramatically increased thresholds for auditory frequency discriminations, and a concurrent visual task dramatically increased thresholds for visual contrast discrimination. However, a concurrent task in a different modality had virtually no effect on primary task thresholds in vision or audition, regardless of whether the tasks were spatially superimposed or separated, and irrespective of task load.

Several previous studies have reported interactions between visual and auditory attentional resources (Driver and Spence, 1994; Spence and Driver, 1996; Spence et al., 2000; Driver and Spence, 2004). However, these studies involved directing attention to different regions of space, whereas we took care to ensure that the spatial

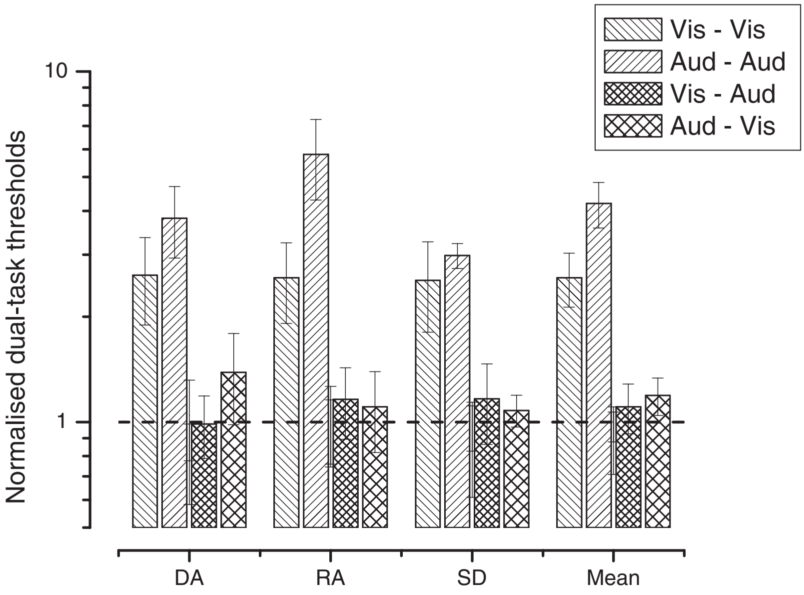

Fig. 9. Threshold performance for three observers (author D.A. and two naïve subjects) for visual and auditory discriminations, all normalized by the single-task threshold. Rightward hatch show visual thresholds for dual tasks with a visual secondary task, leftward hatch auditory thresholds with a visual auditory task. Dense crosshatching shows visual thresholds with an auditory secondary task and sparse crosshatching auditory thresholds with a visual secondary task. The only large effects are for dual tasks in the same modality. Error bars refer to standard errors, calculated by 500 iterations of bootstrapping (Efron and Tibshirani, 1993).

regions stimulated by our visual and auditory stimuli were as similar as possible, and that attention was distributed over the whole field. Furthermore, many of the reported effects were quite small, with *d′* improving from about 0.1 to 0.5 at most (as calculated from their reported error scores). These effects are nearly an order of magnitude less than the intramodal effects we observed. One of our crossmodal conditions showed a very small effect of attention (auditory thresholds measured with visual secondary task), although not the other. However, while statistically significant, the decrement in discriminability caused by the extramodal distractor task was only about 20%, compared to 420% for the intramodal distractor task. So while we cannot totally exclude the existence of crossmodal leakage of attentional limitations, these effects must be considered to be very much secondary compared with the magnitude of intramodal attentional effects.

Although our results are at odds with the conclusions of several recent reports indicating supramodal attentional processes, there is a growing body of evidence indicating independent attentional processes. Our conclusions are in broad agreement with some of the older psychological and human factors literature (Triesman and Davies, 1973; Wickens, 1980), and also agree with those of more recent crossmodal attentional studies using psychophysical and behavioral paradigms quite different to ours (Duncan et al., 1997; Bonnel and Hafter, 1998; Ferlazzo et al., 2002). In addition, a recent transcranial magnetic stimulation experiment that disrupted areas within parietal cortex during visual and somatosensory orienting revealed modality-specific attentional substrates (Chambers et al., 2004), rather than the region being a supramodal attention network (e.g., Macaluso et al., 2002). Other support for our findings comes from recent evidence suggesting that attention is not a unitary phenomenon, but acts at various cortical levels, including early levels of sensory processing and the primary cortical areas of V1 and A1 (Kanwisher and Wojciulik, 2000). Attentional modulation of primary cortices is particularly relevant to our study because the contrast and pitch discrimination tasks used in our experiment are probably mediated by primary cortical areas (Recanzone et al., 1993; Boynton et al., 1999; Zenger-Landolt and Heeger, 2003).

Our results are therefore quite consistent with the notion that each primary cortical area is modulated by its own attentional resources, with very little interaction across modalities. This does not exclude the possibility that attentional effects could also occur at higher levels, after visual and auditory information is combined. Depending on the nature of the task demands, the most sensible strategy might well be to employ a supramodal attentional resource for a given task. For example, speech comprehension in a noisy environment would improve if spatially colocated visual (lip movements) and auditory (speech) signals were combined via a supramodal spatial attention system.

Concluding remarks

Overall, two important general points can be taken from the experiments summarized above. The first is that the Bayesian approach provides a very useful structure for modeling crossmodal interactions. It should be stressed, however, that this approach is largely descriptive, and addresses primarily the issue of how to weight the information from different sources for cue combination. Much work remains to be done to understand how the brain obtains the reliability estimates it needs to use such a framework. Moreover, the model does not address important issues such as the disparity or "conflict" limits beyond which the perceptual system vetoes crossmodal integration.

The second important issue concerns the role of attention. Attention clearly plays an important role in crossmodal interactions, but its nature seems to be more complex than has been previously appreciated. Using low-level stimuli, we found very strong evidence for independent attentional mechanisms for audition and vision. However, important work from other groups shows evidence for supramodal attention in crossmodal tasks. These apparently conflicting results are probably best understood as different aspects of a complex and distributed attentional system that varies in its network organization form one task to another, tailoring itself to optimally perform a particular task. Accordingly, attention will appear to be low level and duplicated unimodally for audiovisual tasks such as we used that are mediated in the primary cortices, but will appear supramodal for tasks involving higher level processes or for tasks where there is uncertainty over which sense should be monitored.

The burgeoning activity in crossmodal research will no doubt shed light on these important matters of attention and Bayesian combination. The flexible nature of attentional processes needs to be understood more fully, and the preattentive aspects of crossmodal interactions need to be specified. As for the Bayesian approach, there is clearly a growing body of evidence highlighting its enormous potential as a model of multisensory combination. Important remaining questions concern the role of knowledge, expectation and attention, and how these factors can be built into a Bayesian framework, most likely by exploiting prior distributions to complement the maximum likelihood combination of ascending sensory signals.

References

Alais, D. and Burr, D. (2003) The "flash-lag" effect occurs in audition and cross-modally. Curr. Biol., 13: 59–63.

Alais, D. and Burr, D. (2004a) No direction-specific bimodal facilitation for audiovisual motion detection. Brain Res. Cogn. Brain Res., 19(2): 185–194.

Alais, D. and Burr, D. (2004b) The ventriloquist effect results from near-optimal bimodal integration. Curr. Biol., 14(3): 257–262.

Alais, D. and Carlile, S. (2005) Synchronizing to real events: subjective audiovisual alignment scales with perceived auditory depth and speed of sound. Proc. Natl. Acad. Sci. USA, 102(6): 2244–2247.

Alais, D. and Weston, E. (2006) Temporal ventriloquism: perceptual shifts in temporal position and improved audiovisual precision predicted by maximum likelihood estimation. J. Vis., 6: 171.

Arrighi, R., Alais, D. and Burr, D. (2005) Neural latencies do not explain the auditory and audio-visual flash-lag effect. Vision Res., 45: 917–2925.

Bald, L., Berrien, F.K., Price, J.B. and Sprague, R.O. (1942) Errors in perceiving the temporal order of auditory and visual stimuli. J. Appl. Psychol., 26: 382–388.

Baldo, M.V. and Klein, S.A. (1995) Extrapolation or attention shift? Nature, 378(6557): 565–566.

Battaglia, P.W., Jacobs, R.A. and Aslin, R.N. (2003) Bayesian integration of visual and auditory signals for spatial localization. J. Opt. Soc. Am. A Opt. Image. Sci. Vis., 20(7): 1391–1397.

Berger, T.D., Martelli, M. and Pelli, D.G. (2003) Flicker flutter: is an illusory event as good as the real thing? J. Vis., 3(6): 406–412.

Bonnel, A.M. and Hafter, E.R. (1998) Divided attention between simultaneous auditory and visual signals. Percept. Psychophys., 60(2): 179–190.

Boynton, G.M., Demb, J.B., Glover, G.H. and Heeger, D.J. (1999) Neuronal basis of contrast discrimination. Vision Res., 39(2): 257–269.

Brefczynski, J.A. and DeYoe, E.A. (1999) A physiological correlate of the 'spotlight' of visual attention. Nat. Neurosci., 2(4): 370–374.

Bremmer, F., Schlack, A., Shah, N.J., Zafiris, O., Kubischik, M., Hoffmann, K., Zilles, K. and Fink, G.R. (2001) Polymodal motion processing in posterior parietal and premotor cortex: a human fMRI study strongly implies equivalencies between humans and monkeys. Neuron, 29(1): 287–296.

Britten, K.H., Newsome, W.T., Shadlen, M.N., Celebrini, S. and Movshon, J.A. (1996) A relationship between behavioral choice and the visual responses of neurons in macaque MT. Vis. Neurosci., 13(1): 87–100.

Britten, K.H., Shadlen, M.N., Newsome, W.T. and Movshon, J.A. (1992) The analysis of visual motion: a comparison of neuronal and psychophysical performance. J. Neurosci., 12(12): 4745–4765.

Bronkhorst, A.W. and Houtgast, T. (1999) Auditory distance perception in rooms. Nature, 397(6719): 517–520.

Burr, D.C., Morrone, M.C. and Banks, M.S. (2006) The ventriloquist effect in time is consistent with optimal combination across senses. J. Vis., 6: 387.

Bushara, K.O., Grafman, J. and Hallett, M. (2001) Neural correlates of auditory-visual stimulus onset asynchrony detection. J. Neurosci., 21(1): 300–304.

Caclin, A., Soto-Faraco, S., Kingstone, A. and Spence, C. (2002) Tactile "capture" of audition. Percept. Psychophys., 64(4): 616–630.

Chambers, C.D., Stokes, M.G. and Mattingley, J.B. (2004) Modality-specific control of strategic spatial attention in parietal cortex. Neuron, 44(6): 925–930.

Clarke, J.J. and Yuille, A.L. (1990) Data Fusion for Sensory Information Processing. Kluwer Academic, Boston.

Colby, C.L., Duhamel, J.R. and Goldberg, M.E. (1993) Ventral intraparietal area of the macaque: anatomic location and visual response properties. J. Neurophysiol., 69(3): 902–914.

Connor, S. (2000) Dumbstruck: A Cultural History of Ventriloquism. Oxford University Press, Oxford.

Corbetta, M. and Shulman, G.L. (2002) Control of goal-directed and stimulus-driven attention in the brain. Nat. Rev. Neurosci., 3(3): 201–215.

Corey, D.P. and Hudspeth, A.J. (1979) Response latency of vertebrate hair cells. Biophys. J., 26(3): 499–506.

Desimone, R. and Duncan, J. (1995) Neural mechanisms of selective visual attention. Ann. Rev. Neurosci., 18: 193–222.

Driver, J. and Spence, C. (1994) Spatial synergies between auditory and visual attention. In: Umiltà, C. and Moscovitch, M. (Eds.) Attention and Performance: Conscious and Nonconscious Information Processing, Vol. 15. MIT Press, Cambridge, MA, pp. 311–331.

Driver, J. and Spence, C. (2004) Crossmodal spatial attention: evidence from human performance. In: Spence, C. and Driver, J. (Eds.), Crossmodal Space and Crossmodal Attention. Oxford University Press, Oxford.

Duncan, J., Martens, S. and Ward, R. (1997) Restricted attentional capacity within but not between sensory modalities. Nature, 387(6635): 808–810.

Eagleman, D.M. and Sejnowski, T.J. (2000) Motion integration and postdiction in visual awareness. Science, 287(5460): 2036–2038.

Efron, B. and Tibshirani, R.J., 1993. An Introduction to the Bootstrap. Monographs on Statistics and Applied Probability, Vol. 57. Chapman & Hall, New York.

Ernst, M.O. and Banks, M.S. (2002) Humans integrate visual and haptic information in a statistically optimal fashion. Nature, 415(6870): 429–433.

Ernst, M.O. and Bülthoff, H.H. (2004) Merging the senses into a robust percept. Trends Cogn. Sci., 8(4): 162–169.

Ferlazzo, F., Couyoumdjian, M., Padovani, T. and Belardinelli, M.O. (2002) Head-centred meridian effect on auditory spatial attention orienting. Q. J. Exp. Psychol. A, 55(3): 937–963.

Gandhi, S.P., Heeger, D.J. and Boynton, G.M. (1999) Spatial attention affects brain activity in human primary visual cortex. Proc. Natl. Acad. Sci. USA, 96(6): 3314–3319.

Ghahramani, Z., Wolpert, D.M. and Jordan, M.I. (1997) Computational models of sensorimotor integration. In: Morasso, P.G. and Sanguineti, V. (Eds.), Self-Organization, Computational Maps and Motor Control. Elsevier Science, Amsterdam, pp. 117–147.

Grady, C.L., Van Meter, J.W., Maisog, J.M., Pietrini, P., Krasuski, J. and Rauschecker, J.P. (1997) Attention-related modulation of activity in primary and secondary auditory cortex. Neuroreport, 8(11): 2511–2516.

Graziano, M.S. (2001) A system of multimodal areas in the primate brain. Neuron, 29(1): 4–6.

Hamlin, A.J. (1895) On the least observable interval between stimuli addressed to disparate senses and to different organs of the same sense. Am. J. Psychol., 6: 564–575.

Hillis, J.M., Ernst, M.O., Banks, M.S. and Landy, M.S. (2002) Combining sensory information: mandatory fusion within, but not between, senses. Science, 298(5598): 1627–1630.

Jacobs, R.A. (1999) Optimal integration of texture and motion cues to depth. Vision Res., 39(21): 3621–3629.

Jacobs, R.A. and Fine, I. (1999) Experience-dependent integration of texture and motion cues to depth. Vision Res., 39(24): 4062–4075.

Jancke, L., Mirzazade, S. and Shah, N.J. (1999) Attention modulates activity in the primary and the secondary auditory cortex: a functional magnetic resonance imaging study in human subjects. Neurosci. Lett., 266(2): 125–128.

Kanwisher, N. and Wojciulik, E. (2000) Visual attention: insights from brain imaging. Nat. Rev. Neurosci., 1(2): 91–100.

King, A.J. and Palmer, A.R. (1985) Integration of visual and auditory information in bimodal neurones in the guinea-pig superior colliculus. Exp. Brain. Res., 60(3): 492–500.

Kopinska, A. and Harris, L.R. (2004) Simultaneity constancy. Perception, 33(9): 1049–1060.

Krekelberg, B. and Lappe, M. (2000) A model of the perceived relative positions of moving objects based upon a slow averaging process. Vision Res., 40(2): 201–215.

Krekelberg, B. and Lappe, M. (2001) Neuronal latencies and the position of moving objects. Trends Neurosci., 24(6): 335–339.

Lamb, T.D. and Pugh Jr., E.N. (1992) A quantitative account of the activation steps involved in phototransduction in amphibian photoreceptors. J. Physiol., 449: 719–758.

Lennie, P. (1981) The physiological basis of variations in visual latency. Vision Res., 21(6): 815–824.

Lewald, J. and Guski, R. (2004) Auditory-visual temporal integration as a function of distance: no compensation for sound-transmission time in human perception. Neurosci. Lett., 357(2): 119–122.

Luck, S.J., Chelazzi, L., Hillyard, S.A. and Desimone, R. (1997) Neural mechanisms of spatial selective attention in areas V1, V2, and V4 of macaque visual cortex. J. Neurophysiol., 77(1): 24–42.

Macaluso, E., Frith, C.D. and Driver, J. (2002) Supramodal effects of covert spatial orienting triggered by visual or tactile events. J. Cogn. Neurosci., 14(3): 389–401.

Mackay, D.M. (1958) Perceptual stability of a stroboscopically lit visual field containing self-luminous objects. Nature, 181(4607): 507–508.

Mateeff, S., Hohnsbein, J. and Noack, T. (1985) Dynamic visual capture: apparent auditory motion induced by a moving visual target. Perception, 14(6): 721–727.

Meyer, G.F. and Wuerger, S.M. (2001) Cross-modal integration of auditory and visual motion signals. Neuroreport, 12(11): 2557–2560.

Mills, A. (1958) On the minimum audible angle. J. Acoust. Soc. Am., 30: 237–246.

Morein-Zamir, S., Soto-Faraco, S. and Kingstone, A. (2003) Auditory capture of vision: examining temporal ventriloquism. Brain Res. Cogn. Brain. Res., 17(1): 154–163.

Moutoussis, K. and Zeki, S. (1997) A direct demonstration of perceptual asynchrony in vision. Proc. R. Soc. Lond. B Biol. Sci., 264(1380): 393–399.

Nijhawan, R. (1994) Motion extrapolation in catching. Nature, 370(6487): 256–257.

Pashler, H.E. (1998) The Psychology of Attention. MIT Press, Cambridge, MA.

Perrott, D. and Saberi, K. (1990) Minimum audible angle thresholds for sources varying in both elevation and azimuth. J. Acoust. Soc. Am., 87: 1728–1731.

Pick, H.L., Warren, D.H. and Hay, J.C. (1969) Sensory conflict in judgements of spatial direction. Percept. Psychophys., 6: 203–205.

Purushothaman, G., Patel, S.S., Bedell, H.E. and Ogmen, H. (1998) Moving ahead through differential visual latency. Nature, 396(6710): 424.

Recanzone, G.H., Schreiner, C.E. and Merzenich, M.M. (1993) Plasticity in the frequency representation of primary auditory cortex following discrimination training in adult owl monkeys. J. Neurosci., 13(1): 87–103.

Shams, L., Kamitani, Y. and Shimojo, S. (2000) Illusions. What you see is what you hear. Nature, 408(6814): 788.

Shams, L., Kamitani, Y. and Shimojo, S. (2002) Visual illusion induced by sound. Brain Res. Cogn. Brain Res., 14(1): 147–152.

Shams, L., Kamitani, Y., Thompson, S. and Shimojo, S. (2001) Sound alters visual evoked potentials in humans. Neuroreport, 12(17): 3849–3852.

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

Spence, C., Ranson, J. and Driver, J. (2000) Cross-modal selective attention: on the difficulty of ignoring sounds at the locus of visual attention. Percept. Psychophys., 62(2): 410–424.

Standage, G.P. and Benevento, L.A. (1983) The organization of connections between the pulvinar and visual area MT in the macaque monkey. Brain. Res., 262(2): 288–294.

Stein, B.E. (1998) Neural mechanisms for synthesizing sensory information and producing adaptive behaviors. Exp. Brain. Res., 123(1-2): 124–135.

Sugita, Y. and Suzuki, Y. (2003) Audiovisual perception: implicit estimation of sound-arrival time. Nature, 421(6926): 911.

Titchener, E.B. (1908) Lectures on the Elementary Psychology of Feeling and Attention. MacMillan, New York.

Triesch, J., Ballard, D.H. and Jacobs, R.A. (2002) Fast temporal dynamics of visual cue integration. Perception, 31(4): 421–434.

Triesman, A.M. and Davies, A. (1973) Divided attention to ear and eye. In: Kornblum, S. (Ed.) Attention and Performance, Vol. 4. Academic Press, New York.

Warren, D.H., Welch, R.B. and McCarthy, T.J. (1981) The role of visual-auditory ‘‘compellingness’’ in the ventriloquism effect: implications for transitivity among the spatial senses. Percept. Psychophys., 30(6): 557–564.

Wickens, C.D., 1980. The structure of attentional resources. In: Attention and Performance, Vol. VIII. Erbaum, Hillsdale, NJ.

Woodruff, P.W., Benson, R.R., Bandettini, P.A., Kwong, K.K., Howard, R.J., Talavage, T., Belliveau, J. and Rosen, B.R. (1996) Modulation of auditory and visual cortex by selective attention is modality-dependent. Neuroreport, 7(12): 1909–1913.

Wuerger, S.M., Hofbauer, M. and Meyer, G.F. (2003) The integration of auditory and visual motion signals at threshold. Percept. Psychophys., 65(8): 1188–1196.

Zahorik, P. (2002) Direct-to-reverberant energy ratio sensitivity. J. Acoust. Soc. Am., 112(5 Pt 1): 2110–2117.

Zenger-Landolt, B. and Heeger, D.J. (2003) Response suppression in v1 agrees with psychophysics of surround masking. J. Neurosci., 23(17): 6884–6893.