Exp Brain Res
DOI 10.1007/s00221-009-1933-z

RESEARCH ARTICLE

Auditory dominance over vision in the perception of interval duration

David Burr · Martin S. Banks · Maria Concetta Morrone

Received: 26 February 2009 / Accepted: 29 June 2009

Abstract The "ventriloquist effect" refers to the fact that vision usually dominates hearing in spatial localization, and this has been shown to be consistent with optimal integration of visual and auditory signals (Alais and Burr in Curr Biol 14(3):257–262, 2004). For temporal localization, however, auditory stimuli often "capture" visual stimuli, in what has become known as "temporal ventriloquism". We examined this quantitatively using a bisection task, confirming that sound does tend to dominate the perceived timing of audio-visual stimuli. The dominance was predicted qualitatively by considering the better temporal localization of audition, but the quantitative fit was less than perfect, with more weight being given to audition than predicted from thresholds. As predicted by optimal cue combination, the temporal localization of audio-visual stimuli was better than for either sense alone.

D. Burr (✉)
Department of Psychology, Università Degli Studi di Firenze,
via S. Nicolò 89, Florence, Italy
e-mail: dave@in.cnr.it

D. Burr
School of Psychology, University of Western Australia,
Nedlands, WA 6009, Australia

M. S. Banks
Vision Science Program, Department of Psychology,
School of Optometry, University of California,
Berkeley, CA, USA

M. C. Morrone
Department of Human Physiological Sciences,
University of Pisa, Via S. Zeno 31, 56100 Pisa, Italy

M. C. Morrone
Scientific Institute Stella Maris, Calambrone, 56018 Pisa, Italy

Keywords Time perception · Vision · Audition · Multi-modal perception

Introduction

When the spatial locations of stimuli specified visually and auditorily are in conflict, vision usually dominates, a well-known fact termed the "ventriloquist effect" (Mateeff et al. 1985; Pick et al. 1969; Radeau 1994; Stekelenburg and Vroomen 2009; Warren et al. 1981). Many explanations have been advanced for the ventriloquist effect, but the most successful is that it is a byproduct of optimal cue combination: if information across senses is weighted according to the statistical reliability of the various sensory signals, vision will determine perceived location because it specifies location more precisely than audition does (Alais and Burr 2004). Strong proof that this is the case is given by the fact that if visual stimuli are blurred, audition dominates. Similar arguments have been made successfully for combination of various forms of multi-modal information (e.g. Clarke and Yuille 1990; Ernst and Banks 2002; Ghahramani et al. 1997).

However, vision does not always dominate hearing. For example, the perceived time of occurrence of a visual stimulus can be influenced by the presentation of an asynchronous auditory stimulus, a phenomenon often called "temporal ventriloquism". Not only is the perceived time of visual stimuli affected (Aschersleben and Bertelson 2003; Fendrich and Corballis 2001) but also the presence of flanking sounds can aid visual discrimination, by increasing their perceived separation (Morein-Zamir et al. 2003; Parise and Spence 2008).

Another example of audition dominating vision is the illusory flash effect of Shams et al. (2000), a seemingly compulsory integration of visual and auditory information

Published online: 14 July 2009

in which audition dominates. When a flashed spot is accompanied by more than one beep, it appears to flash twice, the extra perceived flash being, of course, illusory. This is much like auditory driving, in which the apparent frequency of a flickering visual stimulus can be driven up or down by an accompanying auditory stimulus presented at a different rate (Gebhard and Mowbray 1959; Shipley 1964). Because the auditory system is the most precise sense for temporal judgments, it seems reasonable that it should be the most influential in determining the apparent number of successive stimuli, and also the moment in time when they appear. There are many other examples in which audition seems to affect the interpretation of a visual stimulus (Sekuler and Sekuler 1999; Shams et al. 2001). However, audition does not always dominate over vision in temporal judgments: when sounds are barely above threshold, vision can dominate over audition (Andersen et al. 2004).

One very clear demonstration of sound influencing visual perception comes from a study by Berger et al. (2003). They took advantage of the fact that thresholds for multiple presentations of visual stimuli can be better than those of a single presentation of the same duration (Verghese and Stone 1996). They then associated the visual display of a grating patch with a series of tones (like Shams et al. 2000), and showed not only that the tones cause the visual stimulus to appear as multiple flashes, but also that the apparent multiple flashes actually yielded lower discrimination thresholds, as if they were real multiple presentations of the visual stimulus. This is strong evidence that audition and vision interact in a very real way, and that the apparent increase in the number of perceived flashes is not just a response bias or similar artefact. Further evidence in this direction is the fact that sound can alter visual-evoked potentials in cortical areas as early as V1 (Shams et al. 2001).

To date, most studies of temporal ventriloquism have been qualitative, showing auditory dominance without quantifying the magnitude of the dominance. And none has tested whether the dominance is predictable from optimal cue-combination theory as with visual dominance in the spatial domain (Alais and Burr 2004; Ernst and Banks 2002). In this study, we employ a temporal bisection task to study quantitatively the relative contributions of visual and auditory stimuli to the perceived timing of sensory events. We also test whether the relative contributions of vision and audition can be explained by optimal cue-combination theory.

Methods

Stimuli

The task was temporal bisection as illustrated in Fig. 1. In the two experiments, three stimuli (visual, auditory, or both) were presented in succession for a total duration of 800 ms, and observers were required to indicate by button press whether the middle stimulus appeared closer in time to the first or the third stimulus. Using a procedure similar to that devised by Ernst and Banks (2002), the first and third stimuli were conflict stimuli in which the auditory stimulus was advanced by Δ ms and the visual stimulus delayed by Δ ms ($-60 \leq \Delta \leq 60$ ms): the range was chosen so that the stimulus was always perceived as one rather than two events. The second stimulus was a cue-consistent probe in which the auditory and visual components were presented simultaneously. We adjusted the timing of the second stimulus until it was perceived as bisecting the interval between the first and third stimuli.

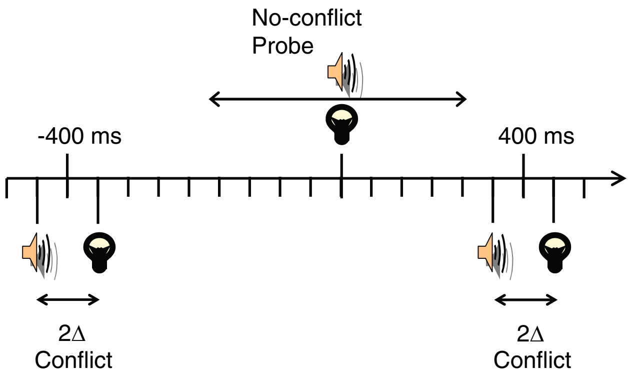


Fig. 1 Illustration of the temporal bisection task. Observers were required to report whether the central stimulus appeared to be nearer in time to the first or last stimulus. In a given condition, the display could comprise only visual flashes, only auditory tones, or both (see "Methods" for complete description of stimuli). For the audio-visual (2-cue) presentations, a conflict between visual and auditory signals was introduced in the first and last stimuli, with the auditory tone leading the visual flash by Δ ms (with Δ varying between ±60 ms in 10 ms steps)

The visual stimuli were 2° diameter disks displayed for 5 ms on a Clinton Monoray monitor equipped with fast-decay DP104 phosphor (decays to 1% in 250 μs). Frame rate was 200 Hz, so visual stimuli consisted of one frame. The luminance of the disks was 20 cd/m^2, and that of the background 10 cd/m^2. Visual stimuli were generated by a VSG V2/5 framestore (Cambridge Research Systems) that operated independently from the PC computer in which it was housed, leaving the computer free to control the auditory stimuli online. Auditory stimuli were pure tones gated with a Gaussian. In the first experiment, the average frequency of the tone was 1,700 Hz, the standard deviation of the Gaussian 10 ms, and the intensity (as source) was 82 dB. In the second experiment, they were 200 Hz, 80 ms standard deviation, and 70 dB. The stimuli in the first experiment were easy to localize in time, and those in the second experiment were more difficult. To minimize pitch cues that could have influenced the task, the frequency of

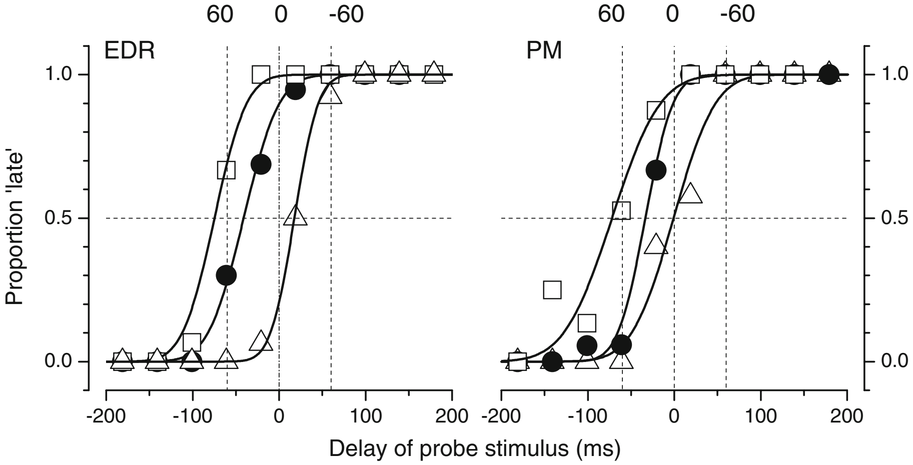


Fig. 2 Example psychometric functions for two subjects with three levels of conflict for Experiment 1 (1,700 Hz stimuli). The data show the proportion of trails judged to be late (closer to second than first marker) as a function of actual time of presentation (where zero refers to midway: see Fig. 1). In all cases the data were well fit by a cumulate gaussian function, whose mean gave an estimate of PSE and standard deviation an estimate of threshold. PSEs shifted systematically with conflict Δ, in the direction of the auditory stimulus. For both subjects, *open square* symbols refer to Δ = −60 ms, *filled circles* to Δ = 0 and *open triangles* to Δ = 60 ms. The *vertical dotted lines* show the temporal position of the auditory standard. For all conflicts, the PSE tended to follow the auditory standard

the tones varied randomly about the mean (Gaussian with a standard deviation of 10% of the mean of 1,700 or 200 Hz). The auditory stimuli were created digitally in Matlab at a sampling rate of 65 kHz and played through two high-quality Yamaha MSP5 loudspeakers mounted adjacent to the display screen, 45 cm from the observer, and ±30 cm from the monitor centre. The two speakers were played in synchrony, so the sound seemed to come from a poorly localized point between them, coinciding roughly with the position of the visual stimulus. Accurate timing of the visual and auditory stimuli was ensured by setting priority in the operating system to maximum during stimulus presentation to thereby avoid interrupts by other processes.

The presentation program waited for a frame-synchronization pulse and then launched the visual and auditory signals. To ensure that the visual and auditory stimuli were always synchronized, a calibration routine was developed that read input from two analogue-to-digital converters (on the VSG) connected to a photocell attached to the monitor and to the speaker input. We inserted a small and constant temporal delay of 3 ms in the auditory stimulus to make the presentation time of the disk coincide exactly with the peak of the auditory temporal Gaussian. Synchrony was checked frequently.

Procedure

Before collecting data, subjects were familiarized with the task in two training sessions of 30 trials each. In those sessions, two-cue, auditory-visual stimuli were presented with no conflict. Subjects indicated after each presentation of three auditory-visual stimuli whether the second appeared earlier or later than the midpoint between the first and third stimuli. We provided feedback during these training sessions, so observers could learn the task and minimize errors in their responses. No feedback was given after the training sessions.

During the experiment proper, 15 different conditions were intermingled within each session: vision only, auditory only, and 13 two-cue conditions with Δ ranging from −60 to 60 ms in 10-ms steps. One session comprised 150 trials (10 for each condition) and was repeated 8–10 times, producing 80–100 points per condition. The time of the probe was varied by independent QUEST routines (Watson and Pelli 1983), perturbed by a Gaussian with a standard deviation of 80 ms (larger than the width of most psychometric curves). The QUEST procedure homed in on the point of subjective equality (PSE): the time offset for which the second stimulus on average appeared to bisect the first and third stimuli. The randomization associated with QUEST ensured that the psychometric function was well sampled for estimating the PSE and slope. It also gave observers a few "easy" trials from time to time. Each experimental session comprised ten trials per condition, and lasted about 10 min. Subjects took breaks at liberty between sessions. Each experiment took ~90–100 min.

Data for each condition were fitted by cumulative Gaussians (see Fig. 2), yielding PSE and threshold estimates from the mean and standard deviation of the best-fitting function, respectively. Standard errors for the PSE and threshold estimates were obtained by bootstrapping (Efron and Tibshirani 1993). All conflict conditions were used to obtain the two-cue threshold estimates: the data were first aligned by subtracting the PSE for each condition, then all fit together (about 1,300 points). There was no systematic variation in the slope of the psychometric functions with size of conflict, so it was reasonable to align and average the curves. Where predictions are accompanied by error bars (for example of the two-cue threshold from single cues), the error bars are standard error of the mean calculated by bootstrap.

Subjects

A total of ten subjects participated in the experiments, the three authors and seven others who were naïve to the goals of the study. All subjects had normal or corrected-to-normal vision and normal hearing. The authors participated only in Experiment 1, another three naïve subjects participated only in Experiment 2, and four (naïve) subjects participated in both experiments.

Results

Experiment 1: high-frequency tones

Figure 2 shows typical psychometric functions for the bisection task for two naïve observers. All three curves are for two-cue presentations, with $\Delta = 0$, -60 or $+60$ ms. All three sets of data are well described by cumulative Gaussians. For the no-conflict condition (filled circles), the 50% point of the curves (the PSE) occur at an offset of about -60 ms. This means that for no-conflict stimuli, the midpoint of the trio is not perceived at the physical midpoint, but ~60 ms earlier. This was a common feature in all the data (see later graphs), and consistent with much other data in the literature, showing that the first interval in a sequence is perceived as longer than the others (Rose and Summers 1995; Tse et al. 2004). It may also reflect the fact that auditory stimuli tend to be perceived earlier than visual stimuli even when they were physical simultaneous (Arrighi et al. 2006; Dixon and Spitz 1980; Summerfield and McGrath 1984), and the auditory stimuli are fundamental in determining perceived timing. More importantly for our purposes, introducing a conflict in the audio-visual presentation shifted the PSE. A positive conflict—with the auditory stimulus preceding the visual stimulus—shifted the curve towards an earlier time. A negative conflict (vision first) shifted the curve the other way. This means that perceived timing follows the auditory more than the visual stimulus.

The effect of conflict on PSE is shown in Fig. 3 for four of the seven observers (three typical, one atypical). In this plot, a slope of +1 indicates total visual dominance, and a slope of -1 total auditory dominance. In all cases, the PSEs varied in an orderly fashion with conflict. Observers EDR, PM, and JM (and three others not shown here) exhibited slopes that approached -1, indicating that the auditory stimulus dominated perceived timing. The slopes were, however, greater than -1 (-0.67 to -0.75), showing that the visual stimuli had an influence on perceived timing. Interestingly, the results for MCM (author) were quite different. She showed far less bias in the no-conflict situation, and very little dependency on conflict, with a best-fitting slope near zero. The upright and inverted triangles show the PSEs for one-cue conditions (respectively vision and audition).

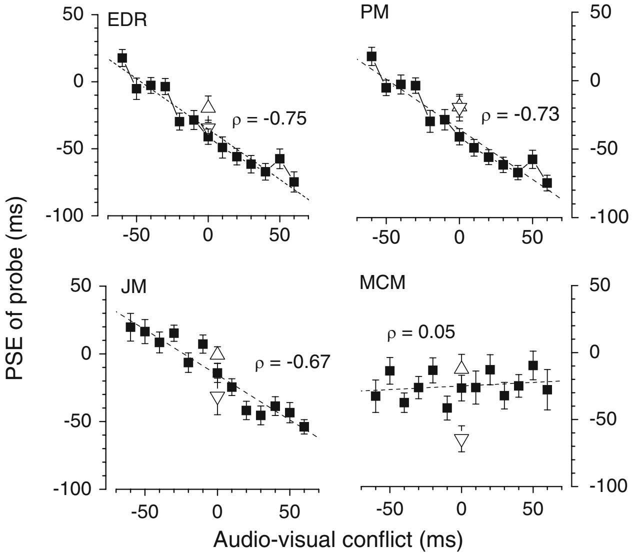


Fig. 3 The effect of audio-visual conflict on PSE for four observers, three typical and one atypical, for Experiment 1 (1,700 Hz stimuli). *Error bars* are calculated by bootstrap (500 reiterations). The *open triangles* refer to 1-cue presentations, *upright triangles* to auditory, *inverted triangles* to vision. The *dashed lines* show the best fitting linear regressions (weighted by standard errors) and the values of ρ near them to the slope of the fit

The dependency of PSEs on the conflict is indexed by the slope of the linear regression of PSE against conflict; this is unaffected by systematic bias that is independent of conflict. The slopes (calculated by regression) for these four observers, together with the other three, are given by the ordinate value of the open squares of Fig. 5. For all observers except the atypical MCM, the slopes are negative, between -0.6 and -1.

Experiment 2: low-frequency tones

In the first experiment, the auditory stimuli were brief (10 ms) tones with a frequency (1,700 Hz) near the peak of the audibility function. Auditory signals clearly dominated the perceived time of arrival, just as visual signals dominated auditory or haptic signals in spatial tasks unless the visual stimulus was degraded (Alais and Burr 2004; Ernst and Banks 2002). We investigated whether degradation of the auditory stimulus would cause a shift toward visual dominance. We degraded the auditory signals as time indicators in two ways: by lowering the frequency from 1,700 to 200 Hz, and by increasing the time constant of the temporal window from 10 to 80 ms, which made the onset less clearly marked.

The results for four observers are shown in Fig. 4. There was considerable variation between observers, but in general the slopes were less negative than those in the first experiment. Again, the results for all observers (seven in total for this condition) are represented by the ordinate values of the filled circles of Fig. 5. While there is considerable spread, these values tend to be to the right of those for the 1,700-Hz tones (filled squares), suggesting a greater dependency on vision when the sound source is degraded. The average slope for 200 Hz was +0.06, compared with −0.60 for 1,700 Hz.

Predictions from single-cue thresholds

As mentioned earlier, several authors (e.g. Clarke and Yuille 1990; Ernst and Banks 2002; Ghahramani et al. 1997) have suggested and demonstrated that multi-modal information may be combined in an optimal way by summing the independent stimulus estimates from each modality according to an appropriate weighting scheme. Assuming the visual and auditory estimates are perturbed by conditionally independent, Gaussian-distributed noise, the weights are inversely proportional to the normalized variance (σ^2) of this noise. We assume that the variance is well estimated from the best-fitting cumulative Gaussian function to the one-cue bisection data. For this experiment, the prediction can be expressed as

$$\hat{T}_{\mathrm{AV}} = w_{\mathrm{A}}\hat{T}_{\mathrm{A}} + w_{\mathrm{V}}\hat{T}_{\mathrm{V}} \tag{1}$$

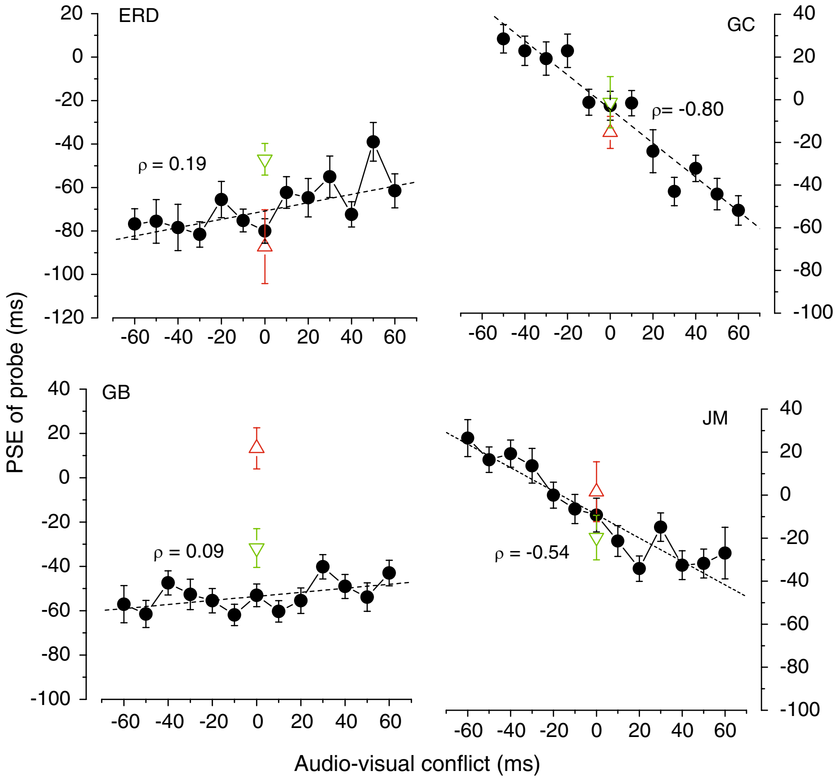
where $\hat{T}_{\mathrm{AV}}$ is the optimal combined estimate of time, $\hat{T}_{\mathrm{A}}$ and $\hat{T}_{\mathrm{V}}$ are the independent estimates for audition and vision, w_{A} and w_{V} are the weights by which the unimodal estimates are scaled. The weights are inversely proportional to the variances σ^2_{A} and σ^2_{V} for audition and vision, normalized to sum to one:

$$w_{\mathrm{A}} = \frac{\sigma_{\mathrm{A}}^{-2}}{\sigma_{\mathrm{V}}^{-2} + \sigma_{\mathrm{A}}^{-2}} \tag{2}$$

$$w_V = \frac{\sigma_V^{-2}}{\sigma_V^{-2} + \sigma_A^{-2}} \tag{3}$$

This model is “optimal” in that it combines unimodal information to produce multimodal estimates with the lowest possible variance (that is, with the greatest reliability: see Clarke and Yuille 1990). In the conflict trials, the position of the visual stimulus was given by Δ and the auditory stimuli by −Δ. Therefore, the combined time estimate $\hat{T}_{\mathrm{AV}}$ should vary with Δ:

Fig. 4 Same as Fig. 3 for Experiment 2, with tones of 200 Hz and gaussian vignette of 80 ms



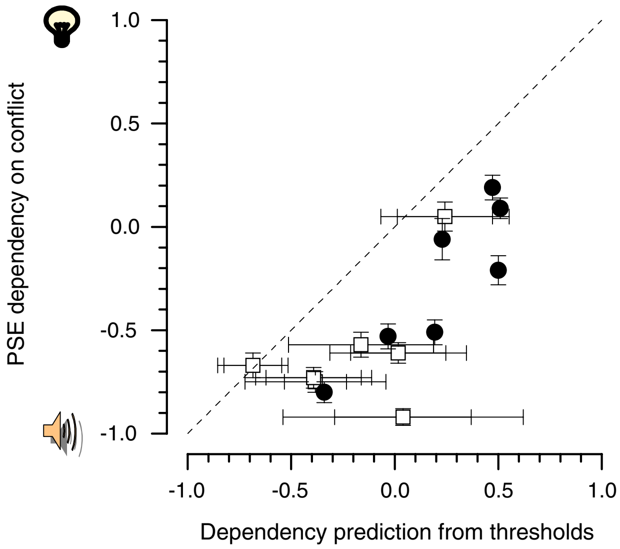


Fig. 5 Measured versus predicted dependency of PSE on audio-visual conflict. The predictions are obtained from the independent thresholds for audio and visual stimuli, from Eq. 7 in text. Error bars for the predictions are obtained by 500 bootstrap reiterations of the equation (resampling the original data from which thresholds were estimated). The measured values (ordinate) are given by the best-fitting regression of curves like those of Figs. 3 and 4, with their associated errors. *Filled square symbols* refer to measurements made with 1,700 Hz tones within 10 ms windows, *open circles* to 200 Hz, 80 ms windows. The *dashed line* represents equality of predicted and measured values. That most points fall clearly below this line suggests the measured dependencies were more negative than the predictions, suggesting that observers gave more weight to auditory signals than predicted by threshold measurements

$$\hat{T}_{AV}(\Delta) = w_V\Delta - w_A\Delta + b \tag{4}$$

where b represents all biases, auditory and visual, that are independent of conflict. As the weights sum to unity, we can eliminate w_V to give

$$\hat{T}_{AV}(\Delta) = (1 - 2w_A)\Delta + b \tag{5}$$

The predicted slope of the function is given by the derivative with respect to Δ:

$$\hat{T}'_{AV}(\Delta) = 1 - 2w_A \tag{6}$$

combining Eqs. 2, 3 and 6 and simplifying

$$\hat{T}'_{AV}(\Delta) = \frac{\sigma_A^2 - \sigma_V^2}{\sigma_V^2 + \sigma_A^2}. \tag{7}$$

σ_A^2 and σ_V^2 can be estimated from the variances of the best-fitting psychometric functions from the auditory and visual one-cue conditions to yield predictions of the slopes of the PSE versus audio-visual conflict functions of Figs. 3 and 4. The estimates for all observers in the two conditions are plotted on the abscissa of Fig. 5, against the measured slope of the two-cue conditions. It is readily apparent that the predictions are not particularly good, for either the 200 or 1,700 Hz conditions. Almost all points lie below the dashed equality line, suggesting that the values of the measured conflict dependency were consistently more negative than the predictions. This means that observers tended to give more weight to the auditory stimulus than predicted by the visual and auditory thresholds, both for the high- and low-frequency tones.

The same data can be plotted in terms of the auditory weights, plotting those calculated from thresholds (Eqs. 2, 3) against those from the slopes of the PSE-conflict curves (rearranging Eq. 6; Fig. 6). Almost all the points lie below equality, reflecting the higher weight predicted from the slope of the one-cue measurements than obtained from the two-cue measurements.

Improvement in thresholds with audio-visual presentations

A very important consequence of optimal combination of information across senses is that it can improve discrimination threshold.

$$\sigma_{VA}^{-2} = \sigma_V^{-2} + \sigma_A^{-2} \tag{8}$$

where σ_{VA}, the threshold of the combined presentation, can never be greater than either the visual or auditory threshold. When visual or auditory variances differ greatly, σ_{AV} is similar to the threshold of the more reliable of the two cues; but when the one-cue thresholds are similar, σ_{AV} will be about $1/\sqrt{2}$ times the values of σ_A and σ_V.

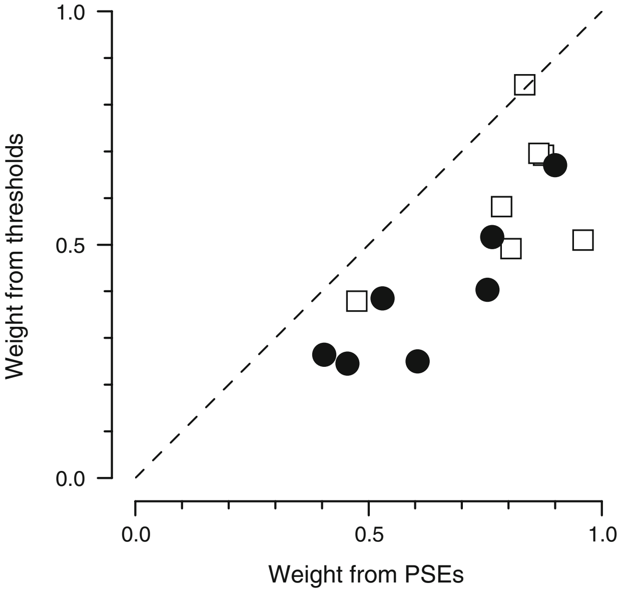


Fig. 6 Same data as in Fig. 5, plotting auditory weights calculated from threshold measurements against weights calculated from the PSE dependency on conflict. Again it is clear that the auditory weighting for the PSE measurements was consistently higher than that predicted by the threshold measurements, for both high and low tones

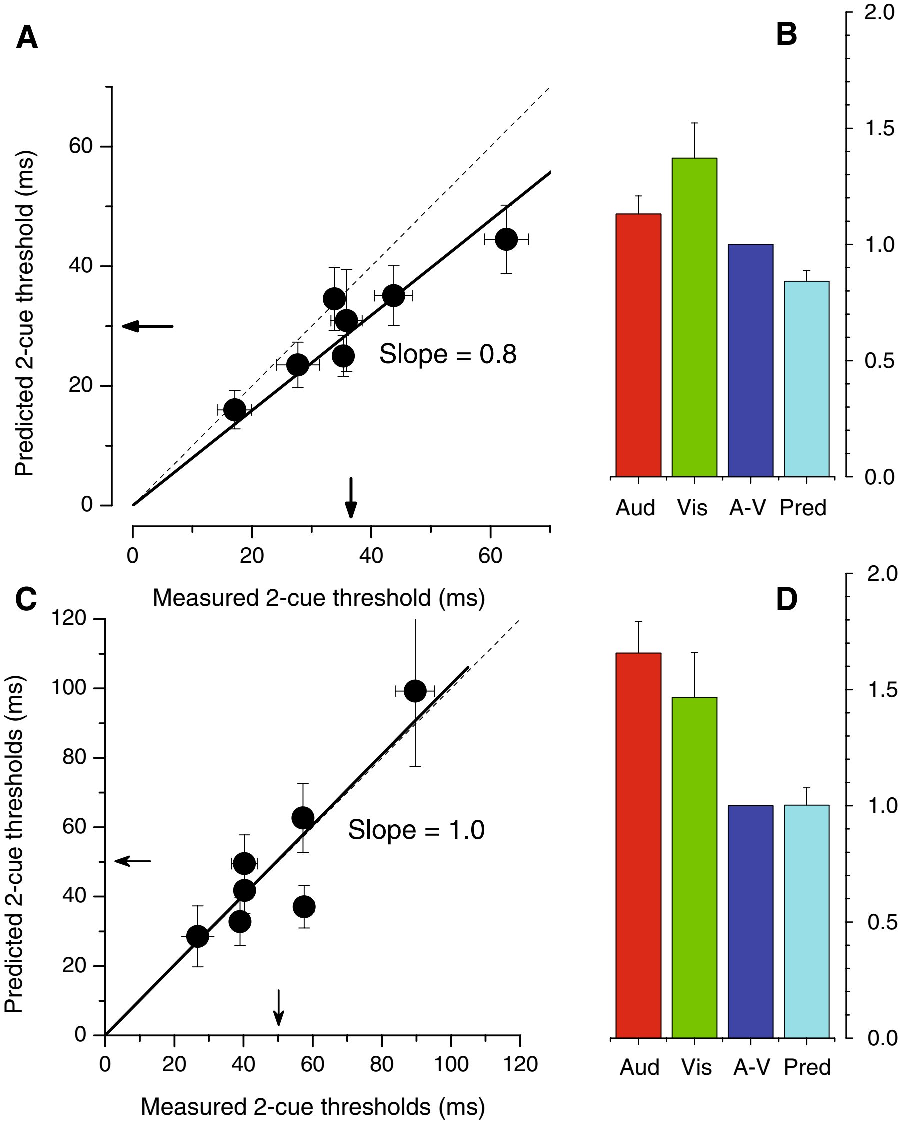


Fig. 7 **a** Predicted 2-cue thresholds (applying Eq. 8 to the visual and auditory threshold measurements) plotted against the actual measurements, for the condition with 1,700 Hz tones. The *arrows* near the axes show the averages. Again, all *error bars* are standard errors from bootstrapping the original data 500 times. The *dashed line* shows the point of equality, the continuous line the best-fitting regression (constrained to pass through zero). The slope of the regression is 0.79 ± 0.04. **b** Mean normalized thresholds for the 1-cue (*red* auditory, *green* visual) and 2-cue conditions (*dark blue*), together with the 1-cue predictions (*light blue*) of the 2-cue thresholds. All thresholds for each individual were first normalized by dividing by their 2-cue threshold before averaging. *Error bars* refer to the standard error of the mean between observers (not taking individual error estimates into account). **c** and **d** same as **a** and **b** except the tones were 200 Hz with gaussian vignette of 80 ms. Note that the prediction is more accurate in this condition

Figure 7a shows the predicted two-cue thresholds for each observer for the 1,700-Hz condition (calculated from the independently measured auditory and visual thresholds from Eq. 8), plotted against the measured two-cue thresholds. The points are reasonably close to, but on average below, the equality line. The best-fitting regression of these points had a slope of 0.79, suggesting that the measured two-cue thresholds were on average slightly greater than the predictions. This is also apparent from the average normalized thresholds in Fig. 7b: the mean predicted two-cue thresholds are about 0.8 times the measured two-cue thresholds. Although the measured thresholds were greater than predicted by optimal cue combination, they were lower than the auditory and visual thresholds. However, the improvement relative to the auditory threshold (the lower of the two) was only marginally significant (paired *t* test, $P = 0.08$).

Figure 7c and d show the results for the 200-Hz tones. Because the one-cue thresholds under these conditions tended to be more similar to each other, giving auditory and visual weights around 0.5 (see ordinate of Fig. 6), the predicted two-cue improvement is greater. Indeed in this case, the prediction was very good. The points of Fig. 7a scatter around the equality line, giving a regression slope of 1.01. The mean of the normalized predictions is almost exactly the same as the measured value (1.002) and clearly less than the best one-cue threshold vision in this case (paired *t* test, $P = 0.008$).

Discussion

This study used a bisection task to examine the effect of auditory and visual stimuli on temporal localization. The

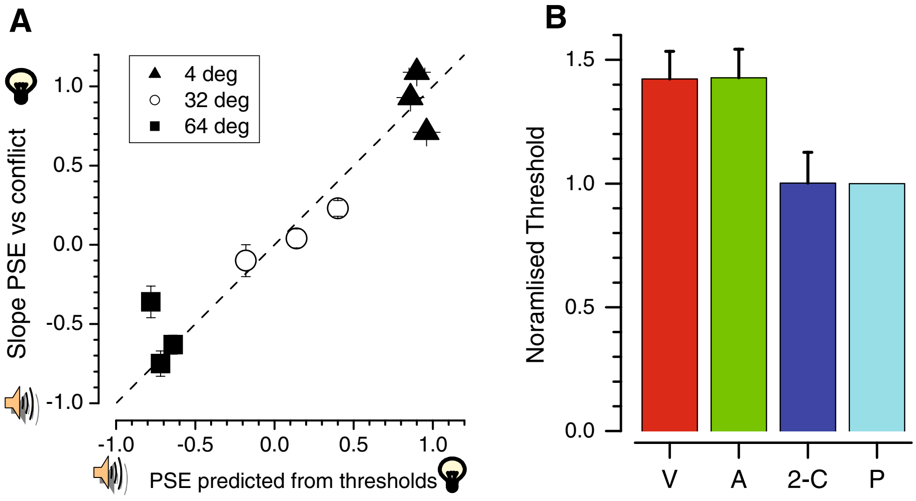


Fig. 8 Summary of the results of the spatial ventriloquist effect (replotted from Alais and Burr 2004), for comparison. **a** Measured versus predicted dependency of PSE on conflict (like Fig. 5). Squares refer to measurements with relatively unblurred stimuli (4° blobs), circles to 32° blobs and *triangles* to 64° blobs. Note the close similarity between predicted and measured dependency (cf Fig. 5); **b** normalized mean thresholds for the 1-cue and 2-cue conditions, together with the 1-cue predictions of the 2-cue thresholds (normalization as in Fig. 7). Again, the prediction is very close to the measured result

results showed that auditory stimuli tend to dominate visual stimuli, but the domination is not total, and varies somewhat from individual to individual. Although the dominance of audition was qualitatively predicted by an optimal model of visuo-auditory combination of temporal information, the prediction was quantitatively imperfect.

Figure 5 summarizes some of the major results. The conflict dependency for most conditions was negative, meaning that the auditory stimulus dominated the perceived temporal position. This was true both for brief tones of optimal frequency (1,700 Hz) and also, albeit to a lesser extent, for 200-Hz tones spread over time within a broad temporal Gaussian envelope. The dominance of audition was seldom total (which would yield a conflict dependency of −1). Interestingly, one subject (MCM) showed roughly equal weighting for vision and audition for the 1,700-Hz stimuli.

Despite the spectacular success of the Bayesian approach in the spatial domain, it does not predict well this particular set of results. Figure 5 shows that the predictions for PSEs from the relative auditory and visual weights, obtained from measuring auditory and visual precision separately, did not match well the measured results, neither for the 1,700-Hz tones nor the 200-Hz tones. By means of comparison, Fig. 8 shows the results for spatial localization, replotted from Alais and Burr (2004). There the dependency on visual or auditory stimuli varied considerably, depending on stimulus blur, but in all cases the results were well predicted by the one-cue precision measurements.

Although the Bayesian approach failed to predict quantitatively the auditory dominance in the PSEs, it worked reasonably well in predicting the improvement in thresholds for the two-cue presentations. For the 200-Hz stimuli, the predicted and observed improvements for two-cue presentations were almost identical. Indeed, the bar graphs of Fig. 7b resemble very closely those of Fig. 8b for spatial localization. For the 1,700-Hz stimuli, the improvement was not exactly as predicted and did not reach statistical significance, but was nevertheless in the right direction.

What could explain the failure of the optimal cue-combination model to predict the auditory dominance of PSEs, given that the model has worked so well in other domains? One possibility is that the estimates of one-cue thresholds are not correct, leading to inappropriate weighting. This could happen, for example, if there were a further non-sensory noise stage related to the bisection that occurred after the fusion of visual and auditory information, related to the judgment of temporal midpoint rather than the localization in time. Imagine that the judgment was over 8 min rather than 800 ms: the bisection noise would clearly swamp any sensory noise, causing a gross overestimation in the sensory noise component. There are several reasons why we think that this explanation cannot account entirely for our results, reasons mainly related to the data obtained with 200-Hz stimuli. First, any additional non-sensory noise source should not vary with stimulus type, and can never be larger than the best threshold performance (in this case auditory and two-cue thresholds at 1,700 Hz). Thus, a level of noise that may be effective for the 1,700-Hz stimuli will be relatively small and ineffective for the 200-Hz stimuli: yet the excessive dominance of audition occurred in both conditions. Second, the effect of a central noise is to overestimate the variance of the more precise sense, tending to make the auditory and visual weights more equal: thus, the predicted conflict dependency will tend towards zero. Again, in the case of the 1,700-Hz tones, this is what happens. However, for the 200-Hz tones, the predictions were for a positive conflict dependency (visual dominance) in five out of seven subjects, which cannot be caused by an additional central noise source. Finally, if non-sensory noise were significant enough to affect the estimates of PSE, then the improvement in two-cue thresholds predicted by Eq. 8 would be violated, because all threshold estimates—visual, auditory and two-cue—would be dominated by the central noise source. Clearly this is not the case for the 200-Hz tones, where the improvement in thresholds almost exactly followed predictions.

Another possibility is that the assumption of Gaussian noise is inappropriate for timing tasks. It is possible that the noise distribution differs significantly from Gaussian, and is possibly not symmetrical, which would affect the predictions. Alternatively, it may be that for timing judgments, weights are not calculated solely from the precision of the individual senses, but auditory information is preferred, possibly because of habit of use, in speech, music, etc. However, this is a difficult notion to test empirically. It should also be pointed out that the optimal cue-combination approach has failed previously in the time domain, in that observers underestimate their temporal variance leading to over-confidence (Mamassian 2008). Why the approach should fail in the temporal, but not the spatial domain, is far from clear.

In conclusion, our results show how auditory stimuli presented at a similar time to visual stimuli can affect the apparent timing of auditory-visual stimuli. When in conflict, sound tends to dominate vision in determining perceived timing, but not totally. The pattern of results was roughly consistent with a model of optimal cue combination, but the quantitative predictions were not accurate. The perceived temporal locations tended to depend more on audition than the threshold measurements suggested they should.

Acknowledgments This research was supported by the Italian Ministry of Universities and Research, EC projects “MEMORY” (FP6-NEST) and “STANIB” (FP7 ERC) and US NIH Research Grant EY-R01-08266 to MSB.

References

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

Andersen TS, Tiippana K, Sams M (2004) Factors influencing audiovisual fission and fusion illusions. Brain Res Cogn Brain Res 21(3):301–308

Arrighi R, Alais D, Burr D (2006) Perceptual synchrony of audiovisual streams for natural and artificial motion sequences. J Vis 6(3):260–268

Aschersleben G, Bertelson P (2003) Temporal ventriloquism: crossmodal interaction on the time dimension. 2. Evidence from sensorimotor synchronization. Int J Psychophysiol 50(1–2):157–163

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

Clarke JJ, Yuille AL (1990) Data fusion for sensory information processing. Kluwer Academic, Boston

Dixon NF, Spitz L (1980) The detection of auditory visual desynchrony. Perception 9(6):719–721

Efron B, Tibshirani RJ (1993) An introduction to the bootstrap. In: Monographs on statistics and applied probability, vol 57. Chapman & Hall, New York

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

Fendrich R, Corballis PM (2001) The temporal cross-capture of audition and vision. Percept Psychophys 63(4):719–725

Gebhard JW, Mowbray GH (1959) On discriminating the rate of visual flicker and auditory flutter. Am J Psychol 72:521–529

Ghahramani Z, Wolpert DM, Jordan MI (1997) Computational models of sensorimotor integration. In: Morasso PG, Sanguineti V (eds) Self-organization, computational maps and motor control. Elsevier, Amsterdam, pp 117–147

Mamassian P (2008) Overconfidence in an objective anticipatory motor task. Psychol Sci 19(6):601–606

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

Morein-Zamir S, Soto-Faraco S, Kingstone A (2003) Auditory capture of vision: examining temporal ventriloquism. Cogn Brain Res 17(15):4–163

Parise C, Spence C (2008) Synesthetic congruency modulates the temporal ventriloquism effect. Neurosci Lett 442(3):257–261

Pick HL, Warren DH, Hay JC (1969) Sensory conflict in judgements of spatial direction. Percept Psychophys 6:203–205

Radeau M (1994) Auditory-visual spatial interaction and modularity. Curr Psychol Cogn 13(1):3–51

Rose D, Summers J (1995) Duration illusions in a train of visual stimuli. Perception 24(10):1177–1187

Sekuler AB, Sekuler R (1999) Collisions between moving visual targets: what controls alternative ways of seeing an ambiguous display? Perception 28(4):415–432

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

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

Shipley T (1964) Auditory flutter-driving of visual flicker. Science 145:1328–1330

Stekelenburg JJ, Vroomen J (2009) Neural correlates of audiovisual motion capture. Exp Brain Res (in press)

Summerfield Q, McGrath M (1984) Detection and resolution of audio-visual incompatibility in the perception of vowels. Q J Exp Psychol A 36(1):51–74

Tse P, Intriligator J, Rivest J, Cavanagh P (2004) Attention and the subjective expansion of time. Percept Psychophys 66:1171–1189

Verghese P, Stone LS (1996) Perceived visual speed constrained by image segmentation. Nature 381(6578):161–163

Warren DH, Welch RB, McCarthy TJ (1981) The role of visual-auditory “compellingness” in the ventriloquism effect: implications for transitivity among the spatial senses. Percept Psychophys 30(6):557–564

Watson AB, Pelli DG (1983) QUEST: a Bayesian adaptive psychometric method. Percept Psychophys 33(2):113–120