

Temporal mechanisms of multimodal binding

David Burr^{1,2,*}, Ottavia Silva³, Guido Marco Cicchini³, Martin S. Banks⁴
and Maria Concetta Morrone^{5,6}

¹Department of Psychology, Università Degli Studi di Firenze, via S. Nicolò 89, Florence 50125, Italy

²School of Psychology, University of Western Australia, Nedlands, Western Australia 6009, Australia

³Faculty of Psychology, Università Vita-Salute San Raffaele, via Olgettina 58, Milan 20132, Italy

⁴Department of Psychology, School of Optometry, Vision Science Program,
University of California, Berkeley, CA 94720, USA

⁵Department of Physiological Sciences, University of Pisa, Via S. Zeno 36, Pisa 56100, Italy

⁶Scientific Institute Stella Maris, Calambrone, Pisa 56018, Italy

The simultaneity of signals from different senses—such as vision and audition—is a useful cue for determining whether those signals arose from one environmental source or from more than one. To understand better the sensory mechanisms for assessing simultaneity, we measured the discrimination thresholds for time intervals marked by auditory, visual or auditory–visual stimuli, as a function of the base interval. For all conditions, both unimodal and cross-modal, the thresholds followed a characteristic ‘dipper function’ in which the lowest thresholds occurred when discriminating against a non-zero interval. The base interval yielding the lowest threshold was roughly equal to the threshold for discriminating asynchronous from synchronous presentations. Those lowest thresholds occurred at approximately 5, 15 and 75 ms for auditory, visual and auditory–visual stimuli, respectively. Thus, the mechanisms mediating performance with cross-modal stimuli are considerably slower than the mechanisms mediating performance within a particular sense. We developed a simple model with temporal filters of different time constants and showed that the model produces discrimination functions similar to the ones we observed in humans. Both for processing within a single sense, and for processing across senses, temporal perception is affected by the properties of temporal filters, the outputs of which are used to estimate time offsets, correlations between signals, and more.

Keywords: time perception; vision; audition; binding; cross-modal

1. INTRODUCTION

One of the most complex tasks for the brain is to combine the information from the five senses into a single perceptual experience. Several studies have shown that the integration of information between senses increases perceptual precision and accuracy (Ernst & Banks 2002; Gepshtein & Banks 2003; Alais & Burr 2004). However, it is crucial that only appropriate information can be integrated because the integration of information from different environmental sources would be generally detrimental.

One cue for when to integrate across modalities could be temporal coincidence: if sensory events (such as a flash and a sound) occur at the same time, there is a good probability that they originated from the same source. But determining simultaneity of external sources is not an easy matter for the brain because the arrival time of neural signals depends on many factors, including variable latencies in sensory transduction and neural transmission, and, for sound, significant physical delays in transmission. Any coincidence detector has to be flexible and adaptable. A good deal of evidence suggests that humans perceive brief auditory and visual events as simultaneous over a moderately wide range of asynchronies. In particular, the system regards auditory–visual

events falling within 50–60 ms of one another as simultaneous (Hirsh & Sherrick 1961; Stein & Meredith 1993; Zampini *et al.* 2003; Arrighi *et al.* 2006). That window is flexible, which is evidenced by the fact that the nervous system takes into account the time the sound takes to travel from its source (Kopinska & Harris 2004; Alais & Carlile 2005). The simultaneity window is also adaptable. Systematic training with asynchronous audio-visual stimuli shifts the time delay at which sounds and flashes are perceived to be simultaneous (Fujisaki *et al.* 2004; see also Vroomen *et al.* 2004). Indeed, artificially delayed visual feedback during a tapping task can distort perceived simultaneity to the extent that when the delay is removed, subjects believe that their actions are anticipating their intentions (Stetson *et al.* 2006).

While the extent of the window of simultaneity for vision and audition (and other senses) has been examined extensively, little is known about the nature of the mechanisms responsible for these tasks. Fujisaki & Nishida (2005) examined synchrony/asynchrony discriminations with periodic stimuli, finding that the discrimination is not possible for frequencies higher than 4 Hz (confirmed by Arrighi *et al.* 2006). They suggested that this limit may reflect a cross-correlation mechanism, computing similarities between auditory and visual streams. Their work further suggested that this cross-correlator does not operate on raw inputs, but correlates salient features extracted by the auditory and visual systems.

* Author and address for correspondence: Department of Psychology, Università Degli Studi di Firenze, via S. Nicolò 89, Florence 50125, Italy (dave@in.cnr.it).

Considering early sensory processing as a cascade of spatial and temporal filters has led to many useful insights, particularly into vision and audition. Here, we use this approach to investigate the filtering properties of auditory–visual synchrony mechanisms.

We measured interval discrimination thresholds where the intervals were marked by visual, auditory and auditory–visual stimuli. Duration thresholds usually follow Weber's law: the required increment in duration is proportional to the base duration (Fraisse 1984; Mauk & Buonomano 2004). While Weber's law is frequently observed in sensory discrimination, there are in fact many important deviations from that behaviour. For example, luminance discrimination departs from Weber's law at low luminances, where the thresholds become independent of luminance (Barlow 1957). More interestingly, many discrimination functions exhibit a 'dipper function', including the functions for discrimination of contrast (Nachmias & Kocher 1970; Pelli 1985), blur (Watt & Morgan 1983; Burr & Morgan 1997) and motion (Simpson & Finsten 1995; Gori *et al.* 2008). Starting with small base values, increment threshold initially decreases with increasing base value reaching the lowest value in the dipper, and then increases monotonically thereafter. For large base values, threshold rises following Weber's law (Nachmias & Kocher 1970; Nachmias & Sansbury 1974; Legge & Foley 1980; Pelli 1985; Foley 1994). Dipper functions have also been observed in visuo-tactile discriminations, where pedestal effects occur between modalities (Arabzadeh *et al.* 2008; Burr *et al.* in press). The generally accepted explanation for the dipper is that it results from a transducer function with an early, threshold-like accelerating nonlinearity (Legge & Foley 1980). Spatio-temporal uncertainty has also been implicated (Pelli 1985), but not strongly supported by the evidence (Legge *et al.* 1987). But whatever the mechanism causing the dipper, it can be modelled by the linear filtering properties of underlying mechanisms, followed by a nonlinearity.

In the study reported here, we measured interval discrimination thresholds as a function of the duration of the base interval for visual, auditory and auditory–visual stimuli. All three conditions produced similar functions with a clear dipper. The dipper occurred at short base intervals for auditory stimuli, slightly longer ones for visual stimuli and much longer ones for auditory–visual stimuli. The results point to the existence of temporal filters with similar properties in the visual and auditory systems, and also in cross-modal integration. These auditory–visual filters could be instrumental in judging auditory–visual simultaneity.

2. MATERIAL AND METHODS

(a) Subjects

Five subjects participated, three females (CL, PB and OS) and two males (GC and LM). The average age was 22.8 years. All had normal or corrected-to-normal vision and normal hearing.

(b) Stimuli and procedure

The experiments were performed in a quiet dark room. Visual stimuli were generated with a VSG 2/5 graphics board (Cambridge Research Systems, Cambridge, UK). Auditory

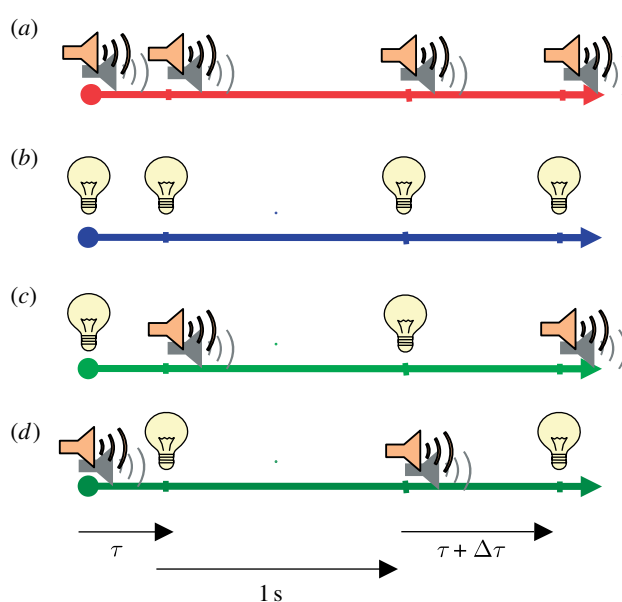


Figure 1. Time course of events in a sample trial. Each trial comprised two intervals delimited either by auditory stimuli ((a) red time line), visual stimuli ((b) blue time line) or by visual then auditory ((c) light green time line) or auditory then visual stimuli ((d) dark green time line). One interval (randomly first or second) had a fixed base duration of τ ms throughout the session, while the other had duration of $\tau + \Delta\tau$ ms, with $\Delta\tau$ varying from trial to trial.

stimuli were generated by the computer soundboard and gated through a switch controlled by the VSG system. This set-up ensured that the visual and auditory stimuli remained in synchrony (calibrated by photocell and microphone). Both systems were controlled by MATLAB running on a PC (Dell, Pentium IV, 2 GHz, 512 Mb RAM). Subjects sat 57 cm from a monitor (Sony, resolution = 464×355 pixels, refresh rate = 160 Hz) that subtended $40 \times 29^\circ$. The speaker was above the monitor.

Visual stimuli were bright circular Gaussian blobs (s.d. = 3.8° , peak luminance = 27.5 cd m^{-2} , CIE coordinates: $x=0.29$, $y=0.31$) presented at the centre of a uniform dark grey background (mean luminance = 1.7 cd m^{-2}) for 13.5 ms (two video frames). The acoustic stimulus was a burst of white noise (sampling rate = 44 100 Hz, intensity = 82 dB(A)) with a Gaussian envelope (s.d. = 5 ms). Stimuli were considered synchronous when the peak of the acoustic envelope was aligned in time with the onset of the second video frame (which corresponds to the centre of the visual stimulus in time). Subjects fixated the centre of the monitor where the visual stimuli appeared. The head position was stabilized with a chin-and-forehead rest.

Two pairs of stimuli were presented on each trial (figure 1) and subjects indicated the pair that seemed longer in duration. In one half of the trial, the stimuli in the pair were separated by a fixed base duration (τ); in the other half, they were separated by the base interval plus an increment ($\tau + \Delta\tau$ ms). There was a 1 s pause between the two halves of the trial, with the base and the base + increment stimuli presented in random order. The base duration varied across sessions from 0 to 400 ms (at a base duration of 0, the single pulse had double intensity). The increment $\Delta\tau$ varied from trial to trial, following an adaptive QUEST routine that homed in on the increment yielding 75 per cent correct (Watson & Pelli 1983). The stimuli in a pair were either both

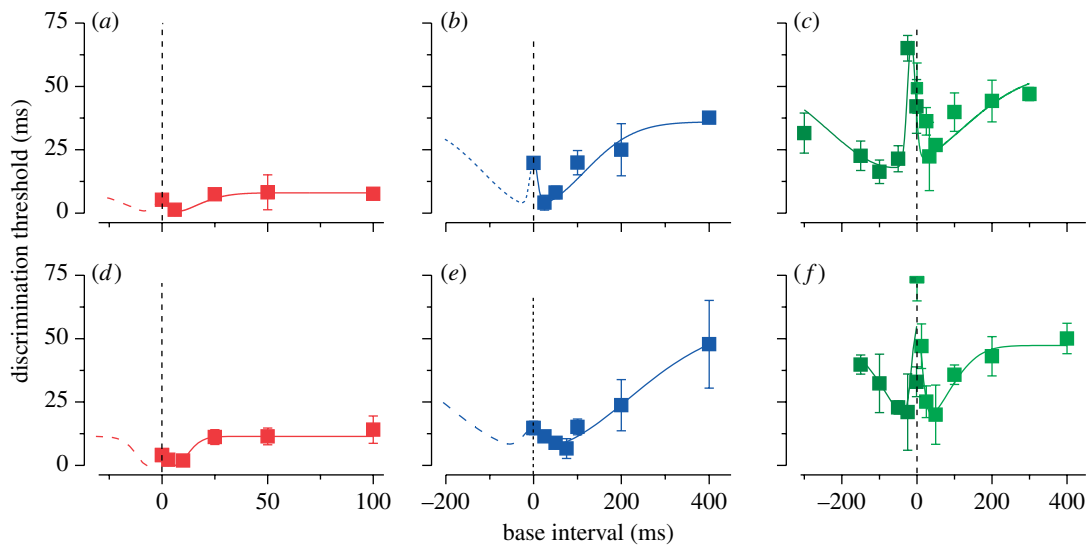


Figure 2. Interval discrimination thresholds for two representative subjects ((a–c) subject GC, (d–f) subject CL) as a function of the base interval τ . Data for the three main conditions are shown in different columns, colour-coded as in figure 1: (a, d) auditory stimuli (red), (b, e) visual stimuli (blue) and (c, f) auditory–visual (green). Negative base intervals represent auditory–visual conditions in which the first stimulus was auditory, positive when the first stimulus was visual. For the unimodal conditions, negative and positive abscissa values are equivalent. Data points represent discrimination thresholds (75% correct). Error bars represent 1 s.e.m. (calculated by bootstrap, 500 repetitions). The continuous curves are the best-fitting difference-of-Gaussian curves, given by

$$\Delta t = A_0 + A_1 e^{-\left(\frac{(\tau - \tau_1)^2}{\sigma_1^2}\right)} - A_2 e^{-\left(\frac{(\tau - \tau_2)^2}{\sigma_2^2}\right)}, \quad (1)$$

where t is interval duration, Δt is the threshold and A_i , τ_i and σ_i are, respectively, the gains, centres and time constants of the Gaussian components. For the auditory and visual conditions, necessarily symmetric, τ_1 and τ_2 were both fixed at 0. The parameters used to fit these and all subjects are shown in table 1.

visual (visual condition), both acoustic (auditory condition) or one visual and one acoustic (auditory–visual condition). The temporal offset in the auditory–visual condition could be negative (sound first) or positive (flash first). Only one condition with one base interval was tested in a given session. Feedback was given after each response. At least 60 trials were run for each condition and base interval. The psychometric data from each condition were fitted with a cumulative Gaussian, and discrimination threshold was given by the median of that best-fitting function. Standard errors were calculated by bootstrap with 500 repetitions.

3. RESULTS

Interval discrimination thresholds as a function of base interval for auditory, visual and auditory–visual stimuli are reported for two representative subjects in figure 2 (the other three showed the same general trend). The auditory–visual data are plotted separately for negative (sound leads) and positive (visual stimulus leads) base durations. The audition-alone and vision-alone data are also plotted with negative and positive base values, but note that the negative data are just a reflection of the ordinate of the positive data. The continuous curves are difference-of-Gaussian functions (five or seven free parameters; equation (1) in legend of figure 2) that best fit the data. A dipper function was observed in all three conditions: as base duration increased, the thresholds first decreased and then increased. The shapes of the functions in the three conditions were similar, but the minimum threshold occurred at a short duration for audition, a slightly longer one for vision and a much longer one for the auditory–visual condition. This suggests that the neural

mechanisms mediating performance with auditory–visual stimuli are more sluggish than the mechanisms that mediate performance with audition- or vision-alone stimuli.

Table 1 summarizes the data for all five subjects, reporting the best-fitting parameters for the difference of Gaussian (equation (1) in legend of figure 2). All subjects behaved similarly, showing a relatively high threshold when the base duration was zero, dropping to a minimum at a base duration as shown in the column ‘dip’. In almost all cases, the value of the dip was shortest for auditory stimuli (average 7 ± 0.3 ms s.e.m.), medium for visual stimuli (average 37 ± 9.5 ms) and longest for auditory–visual stimuli (average positive and negative: 62 ± 6.8 ms). The differences were highly significant on ANOVA ($F_{2,12} = 16.6$, $p = 0.0003$). In the auditory–visual condition, the curves were displaced slightly towards negative values (sound first), with the average midpoint of the two dips being at -20 ms, and the average value of τ_1 (the midpoint of the positive Gaussian) being -12 ms. This suggests that in this experiment there was a slight shift towards sound coming first, suggesting that there were more processing delays associated with sound than vision. However, the effect is very small, and difficult to compare with the previous experiments carried out under different conditions. In the auditory–visual condition subjects reported that they were discriminating apparent simultaneity from non-simultaneity. With the audition- or vision-alone stimuli, the discrimination was between a single and double pulse.

Figure 3 shows, on the abscissa, the duration at which the threshold was lowest (the duration corresponding to the lowest point in the dipper, taken from table 1) for all

Table 1. Values of the best-fitting parameters for equation (1) in the legend of figure 2, used to fit the data of all five subjects. (For the auditory and visual data, τ_1 and τ_2 were fixed at 0 ms (forcing the curves to be symmetrical). The last two columns show the duration at which the curves reached local minima, where dip + means flash first (light green in figures 1–3) and dip – means sound first (dark green).)

	A_0	A_1	τ_1	σ_1	A_2	τ_2	σ_2	dip +	dip –
<i>auditory</i>									
CL	11.4	13.4	—	5.8	20.7	—	11.8	7	—
GC	8	6	—	5.5	9	—	21	7	—
LM	32	3	—	4	8	—	88	6	—
OS	20	3	—	3.5	18	—	56	7	—
PB	18	2.5	—	4.5	16	—	63	8	—
mean	18	6	—	5	14	—	48	7	—
<i>visual</i>									
CL	57	8.5	—	32	50.5	—	305	75	—
GC	36	18	—	12	33	—	161	25	—
LM	44	6	—	13	33	—	182	26	—
OS	36	14	—	15	28	—	146	30	—
PB	46	5	—	15	42	—	164	29	—
mean	44	10	—	17	37	—	192	37	—
<i>bimodal</i>									
CL	47	39	0	19.5	32	0	114	36	–37
GC	56	48	–15	15	38	–60	251	27	–117
LM	45	49	–10	31	24	–93	130	54	–86
OS	134	91	–30	26	93	–20	488	33	–87
PB	109	51	–5	36	92	–69	256	61	–85
mean	78.2	55.6	–12	25.5	55.8	–48.4	248	42.2	–82.4

subjects in all conditions; the circles, squares and triangles represent the individual thresholds and the stars represent the across-subject averages. In the auditory–visual condition, the minimum threshold occurred at longer duration on average when sound-led (negative values on the abscissa) than when vision-led (positive values). Thus, by this metric, the auditory–visual threshold data were shifted towards the sound-first stimuli (shifted leftward in the figure).

The ordinate of the scatter plot shows the minimum increment thresholds for the three conditions. These were lowest for auditory (4.4 ms on average), higher for visual (16 ms) and highest for auditory–visual (62 ms). Indeed, the threshold values were quite similar to the durations at which threshold was minimum. The best-fitting linear regression (on log axes) has a slope of 1.00, is only 0.1 log units below the equality line (dashed diagonal) and accounts for 70 per cent of the variance. This means that the increment thresholds were very similar to the duration at the minimum, as is often found with sensory discrimination thresholds, corroborating the idea that the effect is related to a neuronal threshold (Nachmias & Sansbury 1974).

Figure 4 plots results averaged across subjects on logarithmic coordinates. The change in the duration associated with the dip is evident. The dashed line represents Weber's law (threshold proportional to base duration). The data in the three conditions deviate considerably from one another at short durations, but converge at approximately 400 ms where they approach Weber's law. Presumably, the data would follow Weber's law for yet longer durations.

Our main finding is that interval discrimination functions have the same shape whether the stimuli marking the interval are both auditory, both visual or are cross-modal with audition marking one end of the interval

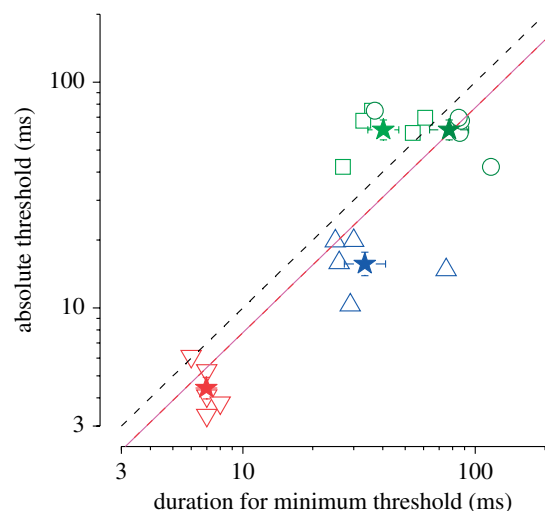


Figure 3. Absolute threshold for duration discrimination (base duration = 0) plotted against the base duration (τ) that produced the lowest thresholds (obtained from the local minima of best-fitting functions as in figure 2, tabulated in table 1). Colour coding is the same as in previous figures: red, auditory; blue, visual; green, auditory–visual (light green for flash first and dark green for sound first). Stars indicate geometrical averages for each condition. The pink line is the best-fitting linear regression (on log–log coordinates); it has a slope of $\rho = 1.00$ and $R^2 = 0.70$. This regression line is approximately 0.1 log units below the equality line (black dashed line), suggesting that the duration at which the minimum threshold is observed is very similar to the absolute thresholds.

and vision marking the other. Thus, the mechanisms that underlie the required temporal measurements have similar properties although they differ in their time constants: shortest for audition and longest for auditory–visual. It is also interesting to note that the Weber fractions for the auditory, visual and auditory–visual conditions are

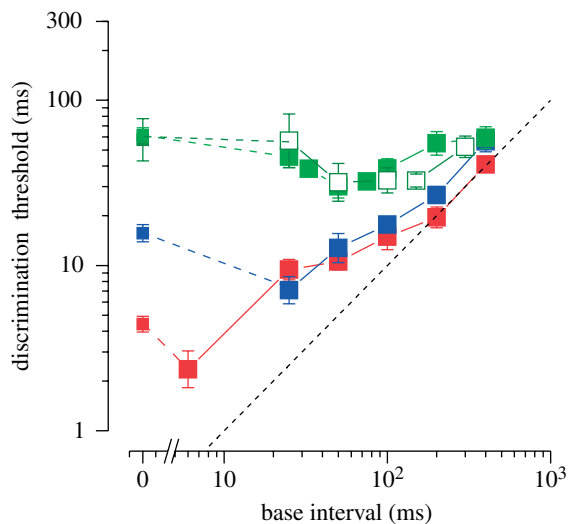


Figure 4. Discrimination thresholds as a function of base interval. The data were averaged over the five subjects (geometric mean) and plotted on logarithmic coordinates. All curves have the same general shape, but the duration associated with minimum threshold varies across conditions. The black dashed line represents unity slope, or Weber's law relationship (red, auditory; blue, visual; green, visual first; white, auditory first).

essentially the same (approx. 10%) for base intervals longer than 200 ms. Thus, performance with long-duration stimulus may be limited by one cross-sensory mechanism that accepts inputs from more than one sense.

(a) Modelling

Although there may be a variety of ways to model these data, we chose to model them with a linear filtering stage, followed by nonlinear feature extraction, and then a decision stage. We did so because this model is biologically plausible and is similar to models used in related areas, such as contrast discrimination (e.g. Legge & Foley 1980). The shape of 'dipper' discrimination functions such as the ones we observed has been linked to nonlinearities in the underlying transducer functions. For example, the contrast-response function for V1 BOLD activation in humans is well predicted by the contrast-discrimination function for each individual subject (Boynton et al. 1999). The prediction is based on the assumption that to discriminate the contrast increment, the neuronal activity should exceed a fixed threshold.¹ The facilitation in the dipper is a consequence of the fact that the nonlinear transducer accelerates at low contrasts, so a small increment can overcome the discrimination threshold. Here, we apply a similar strategy using a linear temporal filter, followed by a nonlinear transducer and then a decision stage.

We assume the filter to be a simple monophasic² temporal impulse response function given by an exponential decay multiplied by a linear increase (equation (2) in the legend of figure 5). The three filters—auditory, visual and auditory–visual—differ only in their time constants of the exponential decay. The filters are linear and time-invariant so that the response to two impulses is simply the sum of the responses to single impulses. Consider the output of the filter presented with two stimuli separated in time (figure 5a–c). When the separation is short (0–50 ms), the response is quite similar to the response to two simultaneous stimuli (figure 5a,b). When the

separation is longer (say 100 ms), the response becomes different in form (figure 5c), with a 'ripple' (deviation from the tangent dashed line) between the two pulses. At large separations, the response becomes clearly biphasic. It is intuitively obvious that any system should be able to detect the difference between figure 5b,c more easily than between figure 5a,b.

The amplitude of the ripple in the response can be quantified by various means, such as measuring the area of the concavity indicated by the blue region. After some experimentation, we chose a simple assumption-free nonlinearity—the sum of the squared response of the filter to the double impulses (Energy)—that decreases as the overlap of the responses to the two individual pulses decreases. Figure 5d shows how the internal representation of the duration of the marked interval $O(\tau)$ varies as a function of the separation of impulses. The dependence of the internal representation on physical duration is clearly not linear, but has the characteristic 'S shape' of transducer functions that are typically invoked to explain dipper functions in domains such as contrast discrimination (Legge & Foley 1980). The function first accelerates to reach a maximum slope for separations similar to the time constant of the filter, then decelerates, eventually becoming constant as the responses to the two impulses become completely separate. Duration discriminations will be best where the slope is at maximum. Thus, discrimination thresholds will first improve as the base interval increases, and then decline as the interval increases further. Note that simply summing the responses without the squaring nonlinearity would produce a linear transducer function, and no dipper. We are not suggesting that the nonlinearity has to arise from a squaring operation, but we do point out that some nonlinearity is needed to create the sigmoidal transducer function and thereby create discrimination data such as those we observed.

We simulated the thresholds from the transducer function by evaluating for each base interval (τ) the minimum increase in interval ($\Delta\tau$) necessary for the internal response $O(\tau)$ to increase by a constant R , set at 10 per cent of the maximum response.

$$O(\tau + \Delta\tau) - O(\tau) = R. \quad (3.1)$$

Figure 5e plots the simulation results along with the average human data. For all three conditions, the model captures the pattern of human thresholds at short and intermediate base durations. The model thresholds initially decrease with increasing duration, creating the characteristic dip, and then increase rapidly. At longer durations, the model behaves differently from humans: its thresholds increase more rapidly than Weber's law. As we said earlier, it is likely that a different kind of mechanism, such as a cascade of filters (Staddon & Higa 1999; Matell & Meck 2004), comes into play at those durations.

Figure 5f shows the temporal impulse response functions that provided the best predictions for the auditory, visual and auditory–visual conditions. All have the same general form, but their time constants differ substantially, being shortest for auditory and longest for auditory–visual. Note that the time constants of the auditory and visual impulse response functions, 9 and 30 ms, respectively, are of an order similar to those obtained by other means (Roufs & Blommaert 1981; Oxenham & Moore 1994; Moore 2003).

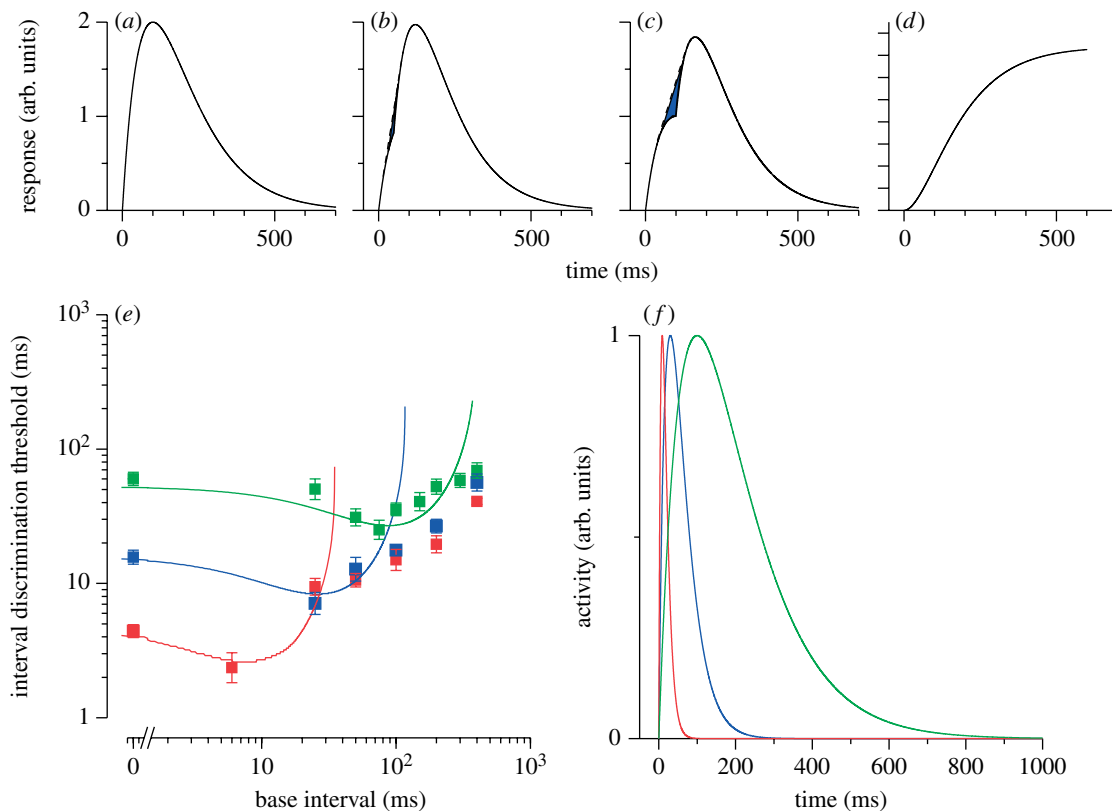


Figure 5. The effect of temporal filtering. (a–d) The outputs to pulses separated by (a) 0, (b) 50 and (c) 100 ms. The filter is a one-lobe, low-pass causal filter given by

$$f(\tau) = \tau \exp(-\tau/\mu), \quad (2)$$

where μ is the time constant of the filter. Examples of the impulse response function (the response of the filter to a single brief pulse) are shown in (f). (a–c) The responses of a filter with $\mu = 100$ ms to double-pulse stimuli with a separation of 1, 50 and 100 ms. The blue-shaded area bounded represents the region of concavity in the curve, with the dashed line bounding this region and the tangent connecting the curve over this concavity. The concavity represents a qualitative difference in the output, clearly more different between separations of 50 and 100 ms than separations 0 and 50 ms. There are many ways to measure this qualitative difference (such as measuring the area of the concavity), but after some experimentation we chose a very simple nonlinearity, the squared response to double stimuli. (d) This measure, expressed as a difference from the response to a fixed separation $\Delta\tau$ and a separation of zero between the pulses. From this internal representation of duration, we calculated for each base duration τ the minimum increase in duration necessary for the response to increase by a threshold value (that we set at 10% of maximum). (e) The results of these simulations (varying μ to give the best fit of the data) against base duration for auditory, visual and bimodal conditions, together with the averaged data. (f) The impulse response functions of the filters that best fit the data. They have time constants μ of 9, 30 and 100 ms, respectively, for the auditory, visual and auditory–visual conditions (red, auditory; blue, visual; green, bimodal).

It should be stressed that none the details of the particular model described above is crucial for its success. The form of filter selected was quite arbitrary, the simplest possible causal low-pass filter. We have simulated our results successfully with various types of filters. Similarly, the way of quantifying the strength of the filter response was arbitrary: many other approaches, such as measuring the depth or area of the ripple in the function (area indicated in figure 5a–c), or the squared difference between the response to a double impulse and to a single impulse (or a prolonged impulse), all gave similar results. What is important is that the response is first filtered, then some nonlinear operator applied. The responses for short impulse separations will then be very similar, irrespective of the means used to measure them.

4. DISCUSSION

Our main finding is that temporal discrimination has similar properties whether the time intervals to be discriminated are marked by two auditory stimuli, two

visual stimuli or an auditory and a visual stimulus. In each case, the discrimination functions exhibit a dipper in which discrimination performance improves with increasing duration and then deteriorates with further increases. The functions differ considerably in temporal scale: the dip occurs at short durations with auditory stimuli; somewhat longer durations with visual stimuli; and much longer durations with auditory–visual stimuli (table 1).

The dipper function has been successfully explained in many conditions by a nonlinear transducer function, be it contrast, blur or variance. We were able to model much of the data in similar fashion. The output of a low-pass filter, followed by a simple nonlinearity (squaring), resulted in the nonlinear transducer function needed to model the data. The model behaves like humans over the range of durations where the dipper was observed. The time constant of the low-pass filter that best fits the data is fastest for audition (9 ms), relatively fast for vision (30 ms) and slowest for the audio-visual judgements (100 ms).

The failure of the model to fit the data at long durations where Weber's law was observed suggests that other mechanisms come into play.

While the idea of temporal (or spatio-temporal) filters is not new to visual or auditory research (de Lange 1954; Robson 1966; Burr 1981), it is perhaps less obvious that a temporal filter exists for auditory–visual judgements. With the unimodal stimuli, both markers pass through the same filter within the visual or the auditory system, transduction filters with properties similar to the cells in visual cortex. The output of those filters can then, after nonlinearities like those we assume, predict the dipper function. When the first and second stimuli come from different modalities (cross-modal condition), they do not both pass through the same transduction system, so these filters could not be responsible. We therefore hypothesize, largely by analogy, that there could exist a higher-stage temporal filter that incorporates visual and auditory inputs, and as a consequence produces a dipper function. If the filter integrates over a longer time interval than is the case with within-modality filters, the dipper in the discrimination data will have a long time constant. Filters with shorter time constants, such as those in the visual and auditory systems, cannot predict this delayed dip.

Perhaps this filter is a part of correlating the streams of incoming auditory and visual stimuli to determine whether the two streams arise from a common environmental source. Analogous behaviour is observed in sound localization and in stereopsis (Stern & Trahiotis 1995; Banks *et al.* 2004). In the case of stereopsis, a sample of one eye's image is correlated with a sample of the other eye's image; the correlation is computed for all relative positions of the two samples in order to find the shift producing the highest correlation. That value is the disparity estimate. A reasonably large spatial sample is required to compute meaningful correlations, and as a consequence the spatial resolution of stereopsis is limited (Banks *et al.* 2004); without relatively large spatial samples, too many high correlations occur, producing false matches. In sound localization, a relatively long temporal sample is required to compute interaural correlations, and the result is that the ability to track changes in source position is sluggish; without using relatively long temporal samples, too many false matches occur (Stern & Trahiotis 1995).

The long temporal constant of the filter examined in the present paper allows integration of information over a reasonable amount of time in order to compute a meaningful correlation between auditory and visual stimuli. The mechanism should also be flexible so that it can be unaffected by constant offsets between auditory and visual stimuli due to differences in physical and neural transduction times. It should also be adaptable in order to change the point of perceived simultaneity after repeated exposure to constant offsets (Fujisaki *et al.* 2004; Stetson *et al.* 2006). One recent result, not obviously consistent with the idea of low-pass temporal filtering, is that Fujisaki & Nishida (2005, 2007) showing that synchrony–asynchrony discrimination is difficult with temporal patterns. However, this can probably be accounted for by assuming a nonlinear temporal 'feature detector' of the sort proposed by Morrone & Burr (1988) and applied successfully to modelling high-order motion detection (Del Viva & Morrone 1998).

Although we chose to model the results within a linear filtering framework, we cannot reject other types of explanation for time-interval discrimination. The standard model for time discrimination is based on ticking clocks and accumulators (e.g. Treisman 1963). To explain Weber's law, which is not predicted by these models, several variations of the clock model have been developed, such as several pacemakers, each ticking at a different rate (Matell & Meck 2004), or hypothesizing that the temporal accumulators have logarithmic sensitivities (Treisman 1963), or relating time estimation to neuronal decay (Staddon & Higa 1999). A model based on a nonlinear accumulator (Treisman 1963) could explain some of the present results. Imagine that the visual or auditory stimulus starts a clock that does not tick uniformly, but follows the nonlinearity seen in the contrast domain: a sigmoidal function, accelerating at low durations and then decelerating at longer ones. A decision stage applied to this nonlinear transducer function would predict the facilitation at the dipper position. However, models of this sort are difficult to evaluate given that at present we do not know the possible biological implementation of a ticking clock or of the accumulator (e.g. Karmarkar & Buonomano 2007).

A phenomenon that is closely related to the issue of perceived simultaneity is 'auditory driving', or the so-called 'temporal ventriloquist effect' (Gebhard & Mowbray 1959; Shipley 1964; Myers *et al.* 1981; Fendrich & Corballis 2001; Morein-Zamir *et al.* 2003), in which the apparent timing of visual stimuli is influenced or 'captured' by auditory stimuli. For example, multiple tone bursts can cause one light flash to appear to be multiple flashes (Shams *et al.* 2000, 2002; Berger *et al.* 2003). This is consistent with the system attempting to perceive stimuli that are close in time as simultaneous. In this conflict situation, the more precise signal determines when the combined stimulus is perceived (Burr & Morrone 2006).

Whatever mechanisms generate the facilitation and dipper functions, it is interesting to consider their possible functional roles. The facilitation (consequence of the accelerating nonlinear transducer) generates a clear boundary between what is seen below or above threshold: i.e. between *perceptual categories*. In the present experiment, visual and auditory stimuli are perceived as simultaneous for a range of non-zero temporal offsets, but when the offsets are larger than approximately 50 ms (depending on the direction of the offset), the stimuli seem non-simultaneous. Thus, there may be two *categories* of perception of auditory–visual stimuli: synchronous and asynchronous. For the single modality data, the category may be between one or two events. The defining characteristic of category perception is that discriminations within the category are more difficult than discriminations between categories (Liberman *et al.* 1957; Studdert-Kennedy *et al.* 1970; Hary & Massaro 1982; Pastore *et al.* 1984; Harnad 1987). This is exactly what we found. Discrimination of durations that were within the 'window for simultaneity' (base intervals near zero) were very difficult: the thresholds were 60–70 ms. However, discriminations relative to a base interval positioned on the border of the category (near absolute threshold for duration discrimination) were much easier, as the two different intervals fell in distinct categories.

The former case is a discrimination within a category (simultaneity) and the latter case across categories. Defining the task in terms of categories does not elucidate the mechanisms responsible for the categorization, but it is a potentially useful way to understand the functional role. Whichever way we choose to look at it, the data suggest that perceived simultaneity holds special importance for our perceptual system and that this is mediated by a filter that combines or correlates visual and auditory signals.

We believe that the results of this study manifest a critical calculation for the nervous system: whether visual and auditory stimuli are derived from one environmental source or more. Without being able to infer whether the current stimulation is due to one or more environmental sources, there would be no point to cue integration. The calculation of source likelihood must work despite the variable neural and physical transmission times that occur with visual and auditory stimulation and processing. Our results describe the properties of the mechanisms that determine temporal coincidence of visual and auditory stimuli. The successful operation of these mechanisms may facilitate many important forms of integration, such as allowing us to attribute spoken words to the appropriate person.

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 M.S.B.

ENDNOTES

¹The threshold does not need to be a 'hard' threshold; an uncertainty model with noise at decision stage would provide the same result.

²Of course, visual-temporal impulse response functions are generally bandpass, but the use of low-pass filters here does not change the argument.

REFERENCES

- Alais, D. & Burr, D. 2004 The ventriloquist effect results from near-optimal bimodal integration. *Curr. Biol.* **14**, 257–262. (doi:10.1016/j.cub.2004.01.029)
- Alais, D. & 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**, 2244–2247. (doi:10.1073/pnas.0407034102)
- Arabzadeh, E., Clifford, C. W. & Harris, J. A. 2008 Vision merges with touch in a purely tactile discrimination. *Psychol. Sci.* **19**, 635–641. (doi:10.1111/j.1467-9280.2008.02134.x)
- Arrighi, R., Alais, D. & Burr, D. 2006 Perceptual synchrony of audiovisual streams for natural and artificial motion sequences. *J. Vis.* **6**, 260–268. (doi:10.1167/6.3.6)
- Banks, M. S., Gepshtein, S. & Landy, M. S. 2004 Why is spatial stereoresolution so low? *J. Neurosci.* **24**, 2077–2089. (doi:10.1523/JNEUROSCI.3852-02.2004)
- Barlow, H. B. 1957 Increment thresholds at low intensities considered as signal/noise discriminations. *J. Physiol.* **136**, 469–488.
- Berger, T. D., Martelli, M. & Pelli, D. G. 2003 Flicker flutter: is an illusory event as good as the real thing? *J. Vis.* **3**, 406–412. (doi:10.1167/3.6.1)
- Boynton, G. M., Demb, J. B., Glover, G. H. & Heeger, D. J. 1999 Neuronal basis of contrast discrimination. *Vision Res.* **39**, 257–269. (doi:10.1016/S0042-6989(98)00113-8)
- Burr, D. C. 1981 Temporal summation of moving images by the human visual system. *Proc. R. Soc. B* **B211**, 321–339. (doi:10.1098/rspb.1981.0010)
- Burr, D. C. & Morgan, M. J. 1997 Motion deblurring in human vision. *Proc. R. Soc. B* **264**, 431–436. (doi:10.1098/rspb.1997.0061)
- Burr, D. & Morrone, C. 2006 Time perception: space-time in the brain. *Curr. Biol.* **16**, R171–R173. (doi:10.1016/j.cub.2006.02.038)
- Burr, D., Gori, M. & Sandini, G. In press. Cross-modal facilitation of visual and haptic motion. *J. Vis.* **9**.
- de Lange, H. 1954 Relationship between critical flicker frequency and a set of low frequency characteristics of the eye. *J. Opt. Soc. Am.* **44**, 380–389. (doi:10.1364/JOSA.44.000380)
- Del Viva, M. M. & Morrone, M. C. 1998 Motion analysis by feature tracking. *Vision Res.* **38**, 3633–3653. (doi:10.1016/S0042-6989(98)00022-4)
- Ernst, M. O. & Banks, M. S. 2002 Humans integrate visual and haptic information in a statistically optimal fashion. *Nature* **415**, 429–433. (doi:10.1038/415429a)
- Fendrich, R. & Corballis, P. M. 2001 The temporal cross-capture of audition and vision. *Percept. Psychophys.* **63**, 719–725.
- Foley, J. M. 1994 Human luminance pattern-vision mechanisms: masking experiments require a new model. *J. Opt. Soc. Am.* **A11**, 1710. (doi:10.1364/JOSAA.11.001710)
- Fraisse, P. 1984 Perception and estimation of time. *Annu. Rev. Psychol.* **35**, 1–36. (doi:10.1146/annurev.ps.35.020184.000245)
- Fujisaki, W. & Nishida, S. 2005 Temporal frequency characteristics of synchrony-asynchrony discrimination of audio-visual signals. *Exp. Brain Res.* **166**, 455–464. (doi:10.1007/s00221-005-2385-8)
- Fujisaki, W. & Nishida, S. 2007 Feature-based processing of audio-visual synchrony perception revealed by random pulse trains. *Vision Res.* **47**, 1075–1093. (doi:10.1016/j.visres.2007.01.021)
- Fujisaki, W., Shimojo, S., Kashino, M. & Nishida, S. 2004 Recalibration of audiovisual simultaneity. *Nat. Neurosci.* **7**, 773–778. (doi:10.1038/nm1268)
- Gebhard, J. W. & Mowbray, G. H. 1959 On discriminating the rate of visual flicker and auditory flutter. *Am. J. Psychol.* **72**, 521–529. (doi:10.2307/1419493)
- Gepshtein, S. & Banks, M. S. 2003 Viewing geometry determines how vision and haptics combine in size perception. *Curr. Biol.* **13**, 483–488. (doi:10.1016/S0960-9822(03)00133-7)
- Gori, M., Sandini, G. & Burr, D. C. 2008 Visual, tactile and visuo-tactile motion discrimination. *J. Vis.* **8**, 173.
- Harnad, S. 1987 Psychophysical and cognitive aspects of categorical perception: a critical overview. In *Categorical perception: the groundwork of cognition* (ed. S. Harnad), pp. 535–565. New York, NY: Cambridge University Press.
- Hary, J. M. & Massaro, D. W. 1982 Categorical results do not imply categorical perception. *Percept. Psychophys.* **32**, 409–418.
- Hirsh, I. J. & Sherrick Jr, C. E. 1961 Perceived order in different sense modalities. *J. Exp. Psychol.* **62**, 423–432. (doi:10.1037/h0045283)
- Karmarkar, U. R. & Buonomano, D. V. 2007 Timing in the absence of clocks: encoding time in neural network states. *Neuron* **53**, 427–438. (doi:10.1016/j.neuron.2007.01.006)
- Kopinska, A. & Harris, L. R. 2004 Simultaneity constancy. *Perception* **33**, 1049–1060. (doi:10.1068/p5169)
- Legge, G. E. & Foley, J. M. 1980 Contrast masking in human vision. *J. Opt. Soc. Am.* **70**, 1458. (doi:10.1364/JOSA.70.001458)

- Legge, G. E., Kersten, D. & Burgess, A. E. 1987 Contrast discrimination in noise. *J. Opt. Soc. Am. A* **4**, 391–404. (doi:10.1364/JOSAA.4.000391)
- Lieberman, A. M., Harris, K. S., Hoffman, H. S. & Griffith, B. C. 1957 The discrimination of speech sounds within and across phoneme boundaries. *J. Exp. Psychol.* **54**, 358–368. (doi:10.1037/h0044417)
- Matell, M. S. & Meck, W. H. 2004 Cortico-striatal circuits and interval timing: coincidence detection of oscillatory processes. *Brain Res. Cogn. Brain Res.* **21**, 139–170. (doi:10.1016/j.cogbrainres.2004.06.012)
- Mauk, M. D. & Buonomano, D. V. 2004 The neural basis of temporal processing. *Annu. Rev. Neurosci.* **27**, 307–340. (doi:10.1146/annurev.neuro.27.070203.144247)
- Moore, B. 2003 *An introduction to the psychology of hearing*. Oxford, UK: Elsevier.
- Morein-Zamir, S., Soto-Faraco, S. & Kingstone, A. 2003 Auditory capture of vision: examining temporal ventriloquism. *Brain Res. Cogn. Brain Res.* **17**, 154–163. (doi:10.1016/S0926-6410(03)00089-2)
- Morrone, M. C. & Burr, D. C. 1988 Feature detection in human vision: a phase-dependent energy model. *Proc. R. Soc. B* **235**, 221–245. (doi:10.1098/rspb.1988.0073)
- Myers, A. K., Cotton, B. & Hilp, H. A. 1981 Matching the rate of concurrent tone bursts and light flashes as a function of flash surround luminance. *Percept. Psychophys.* **30**, 33–38.
- Nachmias, J. & Kocher, E. C. 1970 Visual detection and discrimination of luminance increments. *J. Opt. Soc. Am.* **60**, 382–389. (doi:10.1364/JOSA.60.000382)
- Nachmias, J. & Sansbury, R. V. 1974 Grating contrast: discrimination may be better than detection. *Vision Res.* **14**, 1039–1042. (doi:10.1016/0042-6989(74)90175-8)
- Oxenham, A. J. & Moore, B. C. 1994 Modeling the additivity of nonsimultaneous masking. *Hear Res.* **80**, 105–118. (doi:10.1016/0378-5955(94)90014-0)
- Pastore, R. E., Szczesniul, R., Wielgus, V., Nowikas, K. & Logan, R. 1984 Categorical perception, category boundary effects, and continuous perception: a reply to Hary and Massaro. *Percept. Psychophys.* **35**, 583–588.
- Pelli, D. G. 1985 Uncertainty explains many aspects of visual contrast detection and discrimination. *J. Opt. Soc. Am. A* **2**, 1508–1532. (doi:10.1364/JOSAA.2.001508)
- Robson, J. G. 1966 Spatial and temporal contrast sensitivity functions of the visual system. *J. Opt. Soc. Am.* **56**, 1141–1142. (doi:10.1364/JOSA.56.001141)
- Roufs, J. A. & Blommaert, F. J. 1981 Temporal impulse and step responses of the human eye obtained psychophysically by means of a drift-correcting perturbation technique. *Vision Res.* **21**, 1203–1221. (doi:10.1016/0042-6989(81)90225-X)
- Shams, L., Kamitani, Y. & Shimojo, S. 2000 Illusions. What you see is what you hear. *Nature* **408**, 788. (doi:10.1038/35048669)
- Shams, L., Kamitani, Y. & Shimojo, S. 2002 Visual illusion induced by sound. *Brain Res. Cogn. Brain Res.* **14**, 147–152. (doi:10.1016/S0926-6410(02)00069-1)
- Shipley, T. 1964 Auditory flutter-driving of visual flicker. *Science* **145**, 1328–1330. (doi:10.1126/science.145.3638.1328)
- Simpson, W. A. & Finsten, B. A. 1995 Pedestal effect in visual motion discrimination. *J. Opt. Soc. Am. A Opt. Image Sci. Vis.* **12**, 2555–2563. (doi:10.1364/JOSAA.12.002555)
- Staddon, J. E. & Higa, J. J. 1999 Time and memory: towards a pacemaker-free theory of interval timing. *J. Exp. Anal. Behav.* **71**, 215–251. (doi:10.1901/jeab.1999.71-215)
- Stein, B. E. & Meredith, M. A. 1993 *The merging of the senses*. Cambridge, MA: MIT Press.
- Stern, R. M. & Trahiotis, C. 1995 Models of binaural interaction. In *Handbook of perception and cognition, volume 6: hearing* (ed. B. C. J. Moore), pp. 347–386. New York, NY: Academic Press.
- Stetson, C., Cui, X., Montague, P. R. & Eagleman, D. M. 2006 Motor-sensory recalibration leads to an illusory reversal of action and sensation. *Neuron* **51**, 651–659. (doi:10.1016/j.neuron.2006.08.006)
- Studdert-Kennedy, M., Liberman, A. M., Harris, K. S. & Cooper, F. S. 1970 Theoretical notes. Motor theory of speech perception: a reply to Lane's critical review. *Psychol. Rev.* **77**, 234–249. (doi:10.1037/h0029078)
- Treisman, M. 1963 Temporal discrimination and the indifference interval. Implications for a model of the 'internal clock'. *Psychol. Monogr.* **77**, 1–31.
- Vroomen, J., Keetels, M., de Gelder, B. & Bertelson, P. 2004 Recalibration of temporal order perception by exposure to audio-visual asynchrony. *Brain Res. Cogn. Brain Res.* **22**, 32–35. (doi:10.1016/j.cogbrainres.2004.07.003)
- Watson, A. B. & Pelli, D. G. 1983 QUEST: a Bayesian adaptive psychometric method. *Percept. Psychophys.* **33**, 113–120.
- Watt, R. J. & Morgan, M. J. 1983 The recognition and representation of edge blur: evidence for spatial primitives in human vision. *Vision Res.* **23**, 1465–1477. (doi:10.1016/0042-6989(83)90158-X)
- Zampini, M., Shore, D. I. & Spence, C. 2003 Audiovisual temporal order judgments. *Exp. Brain Res.* **152**, 198–210. (doi:10.1007/s00221-003-1536-z)