BRILL

Multisensory Research 26 (2013) 333–345

MULTISENSORY RESEARCH

brill.com/msr

A Mechanism for Detecting Coincidence of Auditory and Visual Spatial Signals

Emily Orchard-Mills [1,*], **Johahn Leung** [2], **David Burr** [3,4], **Maria Concetta Morrone** [5,6], **Ella Wufong** [7], **Simon Carlile** [2] **and David Alais** [1]

[1] School of Psychology, Brennan MacCallum Building, University of Sydney 2006, New South Wales, Australia

[2] School of Medical Science and The Bosch Institute, University of Sydney 2006, New South Wales, Australia

[3] Department of Neuroscience, University of Florence, via San Salvi 12, 50135 Florence, Italy

[4] Department of Psychology, University of Western Australia, Stirling Hwy., Crawley, Perth, 6907 Western Australia, Australia

[5] Department of Translational Research on New Technologies in Medicine and Surgery, University of Pisa, via San Zeno 31, 56123 Pisa, Italy

[6] Scientific Institute Stella Maris (IRCSS), viale del Tirreno 331, 56018 Calambrone, Pisa, Italy

[7] School of Social Sciences and Psychology, University of Western Sydney, Locked Bag 1797, Penrith 2751, New South Wales, Australia

Received 20 December 2012; accepted 17 June 2013

Abstract

Information about the world is captured by our separate senses, and must be integrated to yield a unified representation. This raises the issue of which signals should be integrated and which should remain separate, as inappropriate integration will lead to misrepresentation and distortions. One strong cue suggesting that separate signals arise from a single source is coincidence, in space and in time. We measured increment thresholds for discriminating spatial intervals defined by pairs of simultaneously presented targets, one flash and one auditory sound, for various separations. We report a 'dipper function', in which thresholds follow a 'U-shaped' curve, with thresholds initially decreasing with spatial interval, and then increasing for larger separations. The presence of a dip in the audiovisual increment-discrimination function is evidence that the auditory and visual signals both input to a common mechanism encoding spatial separation, and a simple filter model with a sigmoidal transduction function simulated the results well. The function of an audiovisual spatial filter may be to detect coincidence, a fundamental cue guiding whether to integrate or segregate.

* To whom correspondence should be addressed. E-mail: emily.orchardmills@sydney.edu.au

DOI:10.1163/22134808-00002425

Keywords
Crossmodal, audiovisual, pedestal, dipper function, spatial representation

1. Introduction

One of the more complex tasks for the brain is to combine information from our senses into a single perceptual experience (Alais *et al.*, 2010; Ernst and Bulthoff, 2004). Integrating information across senses conveys great advantages, as no one sense can capture all kinds of signals from the environment (e.g., light energy, acoustic energy, heat and vibration) and none performs maximally under all conditions. However, it is crucial that only appropriate information be integrated: integrating inappropriately — such as the voice of one speaker with the lip movements of another — can only be detrimental.

One way to determine if a sound and image belong to a single common object is to detect whether they are spatially and temporally coincident. However, this is not a simple task, as the system is necessarily flexible. When simple visual and auditory stimuli are spatially displaced by moderate amounts, they tend to be perceived together, a phenomenon that has become known as the 'ventriloquist effect', as ventriloquism — hearing the ventriloquist's voice emanate from the puppet's mouth — is a common example of the effect. In the laboratory, ventriloquism is usually studied with simple light flashes and sound pulses: when spatially offset, the sound is usually drawn towards the light source (Battaglia *et al.*, 2003; Bertelson and Aschersleben, 1998; Howard and Templeton, 1966; Slutsky and Recanzone, 2001; Welch and Warren, 1980). However, the reverse has also been demonstrated, with audition 'capturing' vision, under conditions where the visual stimuli were degraded sufficiently (Alais and Burr, 2004). It has been shown that audiovisual information about location is combined in a statistically optimal manner, weighting each sensory signal by its reliability (Alais and Burr, 2004), as has been shown previously for other sensory modalities, including touch and vision (Ernst and Banks, 2002; Hillis *et al.*, 2002; Wozny *et al.*, 2008). These studies show that a good deal of spatial mismatch, or conflict, can be tolerated without resulting in the perception of separate objects. There is a clear tendency for the system to fuse auditory and visual stimuli, at least over a certain range.

One of the more interesting interactions between vision and audition is that the visual system can help calibrate the auditory spatial sense. Many studies have shown short-term adaptation to spatially offset auditory and visual inputs (Bermant and Welch, 1976; Held, 1955; Radeau and Bertelson, 1974, 1977, 1978; Recanzone, 1998). Calibration of the auditory spatial representation during development has been demonstrated with young barn owls, reared with distorting prisms, who displayed distorted auditory input, even after the

prisms had been removed (Knudsen and Knudsen, 1985). Recalibration can also occur in adulthood: human subjects who wore lenses that caused compressed vision for a period of days also displayed adaptive changes in sound localisation (Zwiers *et al.*, 2003). Recently, recalibration has been shown to occur at much more rapid timescales, with changes demonstrated after a single trial of audiovisual spatial disparity as brief as 35 ms (Wozny and Shams, 2011a, b). These studies demonstrate the plasticity of the sensory encoding of spatial representations, enabling dynamic updating of the relationship between audition and vision.

Analogous effects also occur with temporal judgments, except that in the temporal domain, audition usually prevails over vision. With the temporal version of the ventriloquism effect, the visual flash tends to be drawn towards the auditory stimulus (Burr *et al.*, 2009a; Fendrich and Corballis, 2001; Hartcher-O'Brien and Alais, 2011; Morein-Zamir *et al.*, 2003). In addition, there are very strong recalibration effects: after hearing asynchronous flash-tone pairings for a period of time, the point of perceived synchrony shifts towards the adaptation offset (Fujisaki *et al.*, 2004; Harrar and Harris, 2008; Navarra *et al.*, 2009; Vroomen *et al.*, 2004).

Burr *et al.* (2009b) measured increment thresholds for discriminating temporal intervals over a range of base intervals. Thresholds for discrimination of intervals defined by two visual markers, two auditory markers or a visual and an auditory marker all revealed 'dipper functions'. Dipper functions describe a pattern of data often observed with increment thresholds whereby starting at small base values, increment thresholds initially decrease with increasing base values before reaching a turning point and then increasing monotonically thereafter for larger base values. Dipper functions have been observed in vision for discrimination of contrast (Nachmias and Kocher, 1970; Pelli, 1985), blur (Burr and Morgan, 1997; Watt and Morgan, 1983), motion (Simpson and Finsten, 1995); in audition for intensity (Hanna *et al.*, 1986; Raab *et al.*, 1963), and across senses for motion (Arabzadeh *et al.*, 2008; Gori *et al.*, 2008, 2011). Burr *et al.* (2009b) argued that the presence of the dipper in the audiovisual data revealed the action of common filters for visual and auditory signals or, equivalently, cross-correlation mechanisms, serving to detect simultaneity. They modelled the results successfully with a linear filtering approach, revealing the action of front-stage audiovisual filters of relatively long time constant. A filter of this sort should facilitate the integration of information from audition and vision, which — based on their proximity in time — are likely to come from a single source.

For visual and auditory stimuli to arise from a common single source, they should be coincident not only in time, but also in space. We therefore studied discrimination of audiovisual spatial intervals, using a similar approach to Burr *et al.* (2009b), measuring increment thresholds for discriminating spa-

tial intervals delimited by one auditory and one visual stimulus, as a function of base interval. The results revealed a 'dipper function'. As in Burr *et al.* (2009b), the presence of a 'dipper function' suggests a mechanism that takes inputs from both vision and audition. This may be functionally useful for signalling spatial co-location across our senses.

2. Methods

2.1. Participants

Six observers (three female) with normal or corrected-to-normal vision and normal hearing participated.

2.2. Stimuli and Apparatus

The experiment was conducted in a darkened anechoic chamber of $4 \times 4 \times 4$ m, with $>99\%$ sound absorption for all frequencies down to 300 Hz (Carlile *et al.*, 1997). Sounds were presented by a speaker (VIFA-D26TG-35) on a robotic arm and the images were projected on an acoustically transparent screen of white muslin. The robotic arm is capable of moving the speaker to any location on an imaginary sphere of 1 m radius around the participant. This arrangement preserves the various localization cues while allowing for the flexible placement of stimuli. Visual stimuli were projected by a Showwx™ Laser Pico Projector, resolution 848×480, refresh rate 60 Hz. It produces no ambient light or noise, ensuring that the room is completely dark, apart from the displayed images. The projector screen (invisible during experiments) was placed 62 cm from the participant, subtending 25° of visual angle.

Participants were seated on a height-adjustable chair and centered in the speaker coordinate system by laser cross hairs. Correct position was facilitated by a chin rest and verified prior to each stimulus presentation with an electromagnetic tracking system (Intersense IC3) attached to the head. Feedback on head position was given to allow participants to correct their alignment, either with LEDs or using the projector.

The visual stimulus was a Gaussian masked circular checkerboard patch 5° wide presented for 100 ms on a black background. The projector output was passed through a series of neutral density filters to reduce the peak luminance of the display to 0.16 cd/m^2 preventing diffused light from illuminating the screen or speaker hoop, so the stimulus was presented without any concurrent visual reference. The auditory stimulus was a 65 dB SPL broadband Gaussian white noise burst of 50 ms duration with a 10 ms raised cosine ramp applied to the onset and offset. After every 10 trials participants were exposed to bright light (33 cd/m^2) for 10 s to prevent dark adaptation, followed by a further period of 10 s of darkness to allow any afterimages to fade before the next trial.

In each interval a pair of perceptually synchronous audiovisual stimuli were displayed. The auditory stimulus was presented after the visual stimulus by a mean of 34 ms (one to three frames, 17–50 ms at 60 Hz). This variation is due to inherent variability in the projector and was confirmed with an oscilloscope. The visual stimulus was presented randomly in one of three possible positions (−5, 0 and 5°), differing for each interval. This prevented participants from making judgements based on the eccentricity of the auditory stimulus alone. The location of the auditory stimulus was varied along the azimuth to produce the desired spatial separation.

2.3. Design and Procedure

A two-interval forced-choice-task was used. In each trial the participants were presented with two intervals in random order, one in which the stimuli were separated by a fixed base interval, the other in which they were separated by the base interval plus an increment, and were required to report which spatial interval appeared larger (see Fig. 1).

Increment thresholds were measured for base intervals ranging from 0 to 15°. The increment was varied by an adaptive QUEST procedure, which converged on the increment giving 75% correct responses. Trials were presented in two blocks, each block containing three base intervals presented in random order. One block contained the base intervals 0, 6 and 12°, the other contained 3, 9 and 15°. The direction of the disparity (left or right) along the azimuth was alternated for each trial. These combinations were used to minimize adap-

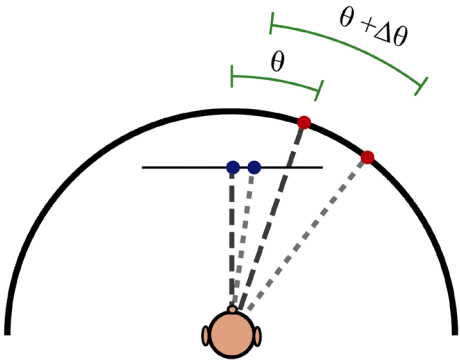

Figure 1. Schematic representation of the task. A trial consists of two intervals presented in random order. In the control interval the targets are separated by a fixed angular distance, the base interval (θ), and in the test interval the targets are separated by this base interval plus an increment ($\theta + \Delta\theta$). The participant responded as to which interval contained the greatest spatial separation. The participant was seated, centered in the speaker hoop, the screen was located inside this hoop. The visual targets are shown on the screen, the auditory targets are shown on the hoop. The increment was varied to measure the increment or just-noticeable difference threshold for each base interval. This figure is published in colour in the online version.

tation, and stereotyping of responses. Presentation of blocks was alternated in each session. The data collected for each base interval was fitted with a Gaussian psychometric function using a constrained maximum-likelihood algorithm as described by Wichmann and Hill (2001). Threshold was taken as the 75% point of the curve.

3. Results

Figure 2 plots increment thresholds as a function of base interval. The thresholds are relatively high for base intervals of zero, then decrease to a minimum

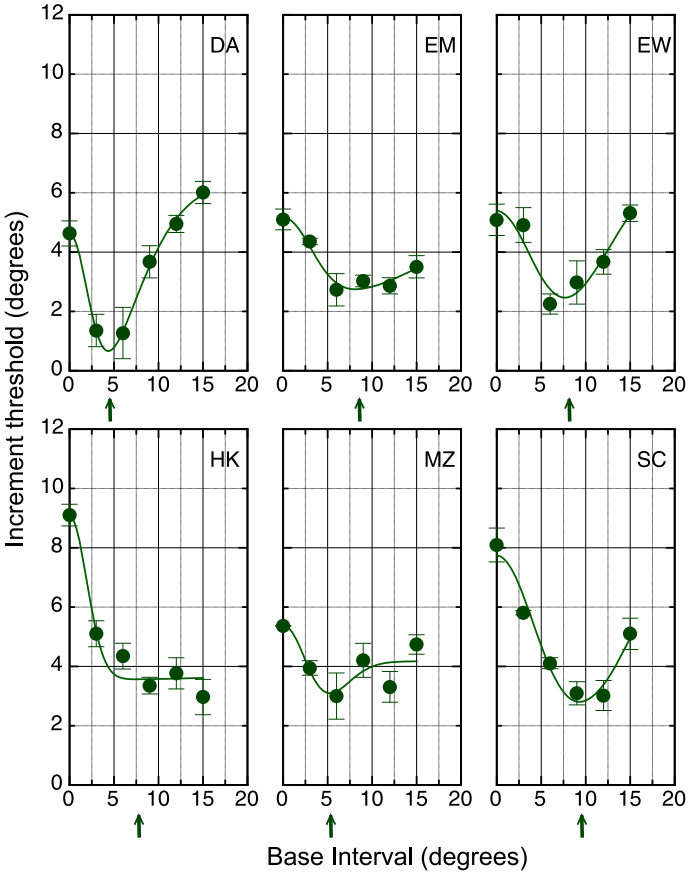

Figure 2. Data from six individuals (initials), with points representing means of increment thresholds (75%) measured in the left and right hemifields. Error bars are standard errors of the mean of the left and right directed intervals. Lines show best fitting difference-of-Gaussian curves to the data (see Table 1 for parameters). Arrows show the interval at which the increment thresholds reached the minimum value. This figure is published in colour in the online version.

for base intervals that are approximately equal to detection threshold (at base-interval zero), then rise again.

A repeated measures analysis of variance (ANOVA) showed a significant effect of base interval, $F(5, 30) = 5.3$, $p < 0.01$. The threshold at a base interval of zero was significantly higher than the thresholds at base intervals of 6, 9 and 12° ($p < 0.05$, Scheffé correction). This confirms that thresholds initially decrease with increasing separation, a key prediction of the 'common audiovisual spatial filter' model.

In order to determine the location of the minimum threshold, the data for each individual subject were fitted with a difference-of-Gaussian curve, as used in Burr *et al.* (2009b), given by

$$\Delta d = A_0 + A_1 e^{-(d^2/\sigma_1^2)} - A_2 e^{-(d^2/\sigma_2^2)},$$

where d is the spatial separation, Δd is the increment threshold, A_0, A_1, A_2 the gains and σ_1, σ_2 are the constants for the Gaussian components. The position of the minimum of this function determined the spatial interval giving the minimum threshold. Best-fitting parameters and the location of the 'dip' for each subject are shown in Table 1 and by arrows on the x-axis of each individual plot in Fig. 2. The group means show that the lowest increment thresholds occurred at a spatial interval of 7.1° (0.8° s.e.m.).

Another key feature of the 'common audiovisual spatial filter' model is that the spatial separation giving the minimum increment threshold should be similar to the increment threshold when there is zero spatial separation. When the mean thresholds were fitted with a difference-of-Gaussian curve, the separation producing the 'dip' and the threshold for zero spatial separation were very similar, 6.5 and 6.1°, respectively. This confirms an important aspect of the 'dipper function', as is discussed below.

Table 1.
Best-fitting parameters for difference-of-Gaussian curves to the individual data (denoted by their initials) shown in Fig. 2. The last column shows the interval at which the increment thresholds reached the minimum value

	A_0	A_1	A_2	σ_1	σ_2	r^2	'dip'
DA	6	7	8	−2	6	0.99	4.4
EM	36415	3	36413	3	1885	0.96	8.2
EW	8	7	10	4	9	0.90	7.7
HK	5	6	1	2	46	0.96	7.6
MZ	4	10	9	3	4	0.70	5.3
SC	225251	8	225251	4	22777	0.96	9.3
Mean	43615	7	43615	2	704	0.91	7.1

4. Discussion

The main finding of this report is that of a 'dipper function' for audiovisual space discrimination, suggesting feature extraction using a spatial filter taking inputs from both vision and audition. Spatial increment thresholds initially improve with increasing spatial separation until a minimum value is reached and thereafter thresholds rise with further increases in spatial separation. In our data, as is typical of dipper functions, the threshold minimum is located at a spatial separation close to the detection threshold.

Following others, we consider processing of audiovisual spatial disparity to involve three stages: linear filtering, non-linear feature extraction and a final decision stage (Burr *et al.*, 2009b; Legge and Foley, 1980). This model relies on the assumption that to detect an increment in spatial disparity — in this case the distance between two stimuli — the output of the filters to the individual stimuli should be discriminably different, with the response to the larger stimulus larger by a fixed threshold quantity. The 'dip' occurs because the feature extraction relies on a non-linear (sigmoidal) transducer. This function has an initial accelerating non-linearity followed by a compressive non-linearity. For small base separations, the slope of the transducer is shallow, so to produce a fixed change in output a large change in separation is required. However, as base separation increases so does the slope of the transducer, with the result being that smaller changes in separation will produce the fixed change in output required. The lowest thresholds will occur on the steepest point of the transducer. After this point the slope of the transducer begins to decrease and the required change in separation will begin to increase. This leads to thresholds that follow a 'U' shape, initially decreasing before turning at a minimum and increasing thereafter.

We suggest a simple model in which following linear filtering at the input level, feature extraction may occur on a spatial representation that takes inputs from both audition and vision. For simplicity we assume a linear, time-invariant Gaussian filter, whose response to two stimuli will be the sum of the individual responses to the stimuli. Figure 3 illustrates the model and the predictions made.

Figure 3a shows the response to two collocated stimuli at the position of the arrow: the individual response is shown by the continuous black curve, and the combined response — in this case a single-peaked Gaussian of twice the amplitude of the single responses — by the dashed line. Figure 3b–d shows the same for two spatially separated stimuli (indicated by arrows). As the separation between the stimuli increases, the combined response becomes broader (c), and eventually double-peaked (d). The key feature of this model is that an increase in separation initially produces little change in the summed response. The difference between (a) and (b) is clearly much less than that of (a) and (d).

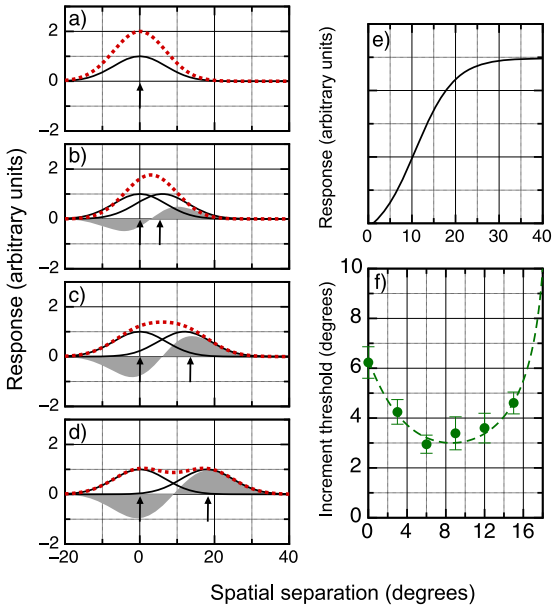

Figure 3. Proposed model of spatial disparity encoding. (a) Two collocated responses (continuous black line) which sum to a single peak of double size (dashed line). (b–d) Two responses (continuous black lines) separated by increasing distance. Initially the sum (dashed line) forms to a single peak, (b), this becomes flattened in (c) and double-peaked in (d). The difference of the separated from collocated responses is shown by the grey shaded areas. Initially, increasing separations result in little change in the summed response, however at a critical point, rapid changes in the summed output occur. It is clear that the difference between (a) and (d) is greater than between (a) and (b). The sum of the squared difference (integrating the grey area), is shown in (e), with a sigmoidal pattern of accelerating non-linearity followed by compressive non-linearity. The increase in separation required to increase the response of the function shown in (e) by a fixed amount (set at 18% of maximum output) for each base separation is shown by the dashed line in (f). The predicted thresholds initially decrease, reach a minimum at a separation around 8.7°, and then increase with further separation. The data points in (f) are mean thresholds averaged across participants with error bars representing standard errors of the mean. This figure is published in colour in the online version.

There are a number of possible ways the system may estimate the distance of two stimuli, in order to estimate which pair had the greater separation. Here we simply subtract the response to one stimulus from that of the other, giving us a difference function (indicated by the grey shaded areas of Fig. 3b–d). We integrate the squared response for each stimulus separation, and plot the result in Fig. 3e. The resulting function has the typical sigmoidal pattern of accelerating non-linearity followed by compressive non-linearity, common to many visual transducer functions (Legge and Foley, 1980).

In modelling the data, thresholds were simulated from the modelled transducer for the range of base intervals tested by determining the minimum

increase in separation required to change the response of the transducer by a constant amount, which we have defined as a percentage of the maximum output of the transducer. We allowed both this 'response increment' constant and the width of the Gaussian to vary to determine the best fitting parameters for the audiovisual thresholds. Figure 3f shows the predicted thresholds (dashed line) from this model and the mean thresholds from the audiovisual data (filled circles). The best-fitting parameters were a Gaussian width of 7.1° and a response increment of 18%. The predicted thresholds fit the pattern of observed data well ($r^2 = 0.96$). The model predicts thresholds that initially decrease with separation, reach a minimum at a threshold of 8.7°, and then increase with further separation.

The simplest alternative model is one where the auditory and visual signals do not combine in a cross-sensory spatial filter but remain separate at early stages, combining only at a later 'decision' stage if an audiovisual spatial judgement is required. This alternative model predicts the lowest threshold for a spatial separation of zero, and a log-linear increase in spatial increment thresholds as base separation increases (that is, a Weber function across the entire range of spatial separation). This monotonic linear prediction is very different from the data we report and it cannot capture its characteristic non-linearity in which thresholds initially drop with spatial separation before rising again at greater separations. The best fit that this late combination model could provide with our observed thresholds was poor, $r^2 = 0.2$. Our audiovisual filter model therefore better describes the data, qualitatively and quantitatively.

Another approach that has been previously suggested for visual spatial interval discrimination is the ratio of the width between the peaks and the width of the overall sum (Solomon, 2009). Using this approach, when the response to the two locations sum to a single peak, in Fig. 3a–c, the distance between the peaks is effectively zero. After the summed response becomes bimodal, this ratio rapidly increases with compressive non-linearity. Though this does predict thresholds that initially decrease with spatial separation before rising again, it produces a near-zero threshold at the dip, a feature not present in the thresholds measured here, and a dip with an implausible V-shaped turning point. A consistent point, however, between this model and the dipper model we propose is that both interpret the 'dipper function' as a result of a 'common audiovisual spatial filter'. We do not suggest that these are the only approaches that can model the 'dip' in the thresholds, however a common filter, taking inputs from both vision and audition, and a transformation resulting in a non-linear transducer are required.

Finding a 'dipper function' for the audiovisual thresholds suggests feature extraction using a filter that takes inputs from both vision and audition. This filter may be responsible for cross-correlating auditory and visual inputs to determine whether two inputs originate from a single environmental source.

This cross-correlation process is analogous to that of the audiovisual temporal filters proposed by Burr *et al.* (2009b), and to models of binaural processing (Stern and Trahiotis, 1995) and of spatial stereovision (Banks *et al.*, 2004). Consistent with the models for these processes, a balance must be achieved between sensitivity to maximise correct binding and specificity to reject false matches.

In summary, we have reported thresholds for a spatial increment discrimination task using spatial intervals defined by pairs of concurrently presented visual and auditory targets. When plotted as a function of base interval, increment thresholds do not rise monotonically but instead form a 'dipper function' where thresholds initially decline and reach a minimum at non-zero base intervals before rising thereafter. The presence of a 'dipper' signature suggests the existence of a spatial filter that takes inputs from both vision and audition. The observed thresholds were successfully modelled using a three-stage process, comprising linear filtering, non-linear feature extraction and a decision stage. The mechanism suggested here could be used by our sensory systems to perform the important task of determining whether auditory and visual spatial signals are coincident or not, analogous to a recent model for detecting temporal coincidence proposed by Burr *et al.* (2009b).

References

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

Alais, D., Newell, F. N. and Mamassian, P. (2010). Multisensory processing in review: From physiology to behaviour, *Seeing Perceiving* **23**, 3–38.

Arabzadeh, E., Clifford, C. W. and Harris, J. A. (2008). Vision merges with touch in a purely tactile discrimination, *Psychol. Sci.* **19**, 635–641.

Banks, M. S., Gepshtein, S. and Landy, M. S. (2004). Why is spatial stereoresolution so low? *J. Neurosci.* **24**, 2077–2089.

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* **20**, 1391–1397.

Bermant, R. I. and Welch, R. B. (1976). Effect of degree of separation of visual–auditory stimulus and eye position upon spatial interaction of vision and audition, *Percept. Motor Skill.* **42**, 487–493.

Bertelson, P. and Aschersleben, G. (1998). Automatic visual bias of perceived auditory location, *Psychon. B. Rev.* **5**, 472–489.

Burr, D., Banks, M. S. and Morrone, M. C. (2009a). Auditory dominance over vision in the perception of interval duration, *Exp. Brain Res.* **198**, 49–57.

Burr, D. C. and Morgan, M. J. (1997). Motion deblurring in human vision, *Proc. Biol. Sci.* **264**, 431–436.

Burr, D., Silva, O., Cicchini, G. M., Banks, M. S. and Morrone, M. C. (2009b). Temporal mechanisms of multimodal binding, *Proc. Biol. Sci.* **276**, 1761–1769.

Carlile, S., Leong, P. and Hyams, S. (1997). The nature and distribution of errors in sound localization by human listeners, *Hearing Res.* **114**, 179–196.

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

Ernst, M. O. and Bulthoff, H. H. (2004). Merging the senses into a robust percept, *Tr. Cogn. Sci.* **8**, 162–169.

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

Fujisaki, W., Shimojo, S., Kashino, M. and Nishida, S. (2004). Recalibration of audiovisual simultaneity, *Nat. Neurosci.* **7**, 773–778.

Gori, M., Mazzilli, G., Sandini, G. and Burr, D. (2011). Cross-sensory facilitation reveals neural interactions between visual and tactile motion in humans, *Front. Psychol.* **2**, 55.

Gori, M., Sandini, G. and Burr, D. C. (2008). Visual, tactile and visuo-tactile motion discrimination, *J. Vision* **8**, 173.

Hanna, T. E., von Gierke, S. M. and Green, D. M. (1986). Detection and intensity discrimination of a sinusoid, *J. Acoust. Soc. Am.* **80**, 1335–1340.

Harrar, V. and Harris, L. R. (2008). The effect of exposure to asynchronous audio, visual and tactile stimulus combinations on the perception of simultaneity, *Exp. Brain Res.* **186**, 517–524.

Hartcher-O'Brien, J. and Alais, D. (2011). Temporal ventriloquism in a purely temporal context, *J. Exp. Psychol. Human* **37**, 1383–1395.

Held, R. (1955). Shifts in binaural localization after prolonged exposures to atypical combinations of stimuli, *Am. J. Psychol.* **68**, 526–548.

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**, 1627–1630.

Howard, I. P. and Templeton, W. B. (1966). *Human Spatial Orientation*. Wiley, New York, NY.

Knudsen, E. I. and Knudsen, P. F. (1985). Vision guides the adjustment of auditory localization in young barn owls, *Science* **230**, 545–548.

Legge, G. E. and Foley, J. M. (1980). Contrast masking in human vision, *J. Opt. Soc. Am.* **70**, 1458–1471.

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

Nachmias, J. and Kocher, E. C. (1970). Visual detection and discrimination of luminance increments, *J. Opt. Soc. Am.* **60**, 382–389.

Navarra, J., Hartcher-O'Brien, J., Piazza, E. and Spence, C. (2009). Adaptation to audiovisual asynchrony modulates the speeded detection of sound, *Proc. Natl Acad. Sci. USA* **106**, 9169–9173.

Pelli, D. G. (1985). Uncertainty explains many aspects of visual contrast detection and discrimination, *J. Opt. Soc. Am. A* **2**, 1508–1532.

Raab, D. H., Osman, E. and Rich, E. (1963). Intensity discrimination, the "pedestal" effect and 'negative masking' with white-noise stimuli, *J. Acoust. Soc. Am.* **35**, 1053.

Radeau, M. and Bertelson, P. (1974). The after-effects of ventriloquism, *Q. J. Exp. Psychol.* **26**(1), 63–71.

Radeau, M. and Bertelson, P. (1977). Adaptation to auditory–visual discordance and ventriloquism in semirealistic situations, *Percept. Psychophys.* **22**, 137–146.

Radeau, M. and Bertelson, P. (1978). Cognitive factors and adaptation to auditory–visual discordance, *Percept. Psychophys.* **23**, 341–343.

Recanzone, G. H. (1998). Rapidly induced auditory plasticity: The ventriloquism aftereffect, *Proc. Natl Acad. Sci. USA* **95**, 869–875.

Simpson, W. A. and Finsten, B. A. (1995). Pedestal effect in visual motion discrimination, *J. Opt. Soc. Am. A* **12**, 2555–2563.

Slutsky, D. A. and Recanzone, G. H. (2001). Temporal and spatial dependency of the ventriloquism effect, *Neuroreport* **12**, 7–10.

Solomon, J. A. (2009). The history of dipper functions, *Atten. Percept. Psycho.* **71**, 435–443.

Stern, R. M. and Trahiotis, C. (1995). Models of binaural interaction, in: *Handbook of Perception and Cognition*, B. C. J. Moore (Ed.), Vol. 6, Hearing, pp. 347–386. Academic Press, New York, NY, USA.

Vroomen, J., Keetels, M., de Gelder, B. and Bertelson, P. (2004). Recalibration of temporal order perception by exposure to audio-visual asynchrony, *Cogn. Brain Res.* **22**, 32–35.

Watt, R. J. and Morgan, M. J. (1983). The recognition and representation of edge blur: Evidence for spatial primitives in human vision, *Vision Res.* **23**, 1465–1477.

Welch, R. B. and Warren, D. H. (1980). Immediate perceptual response to intersensory discrepancy, *Psychol. Bull.* **88**, 638–667.

Wichmann, F. A. and Hill, N. J. (2001). The psychometric function: I. Fitting, sampling and goodness of fit, *Percept. Psychophys.* **63**, 1293–1313.

Wozny, D. R., Beierholm, U. R. and Shams, L. (2008). Human trimodal perception follows optimal statistical inference, *J. Vision* **8**, 24.

Wozny, D. R. and Shams, L. (2011a). Recalibration of auditory space following milliseconds of cross-modal discrepancy, *J. Neurosci.* **31**, 4607–4612.

Wozny, D. R. and Shams, L. (2011b). Frontiers: Computational characterization of visually induced auditory spatial adaptation, *Front. Integr. Neurosci.* **5**, 75.

Zwiers, M. P., Van Opstal, A. J. and Paige, G. D. (2003). Plasticity in human sound localization induced by compressed spatial vision, *Nat. Neurosci.* **6**, 175–181.