

Visual search for a target changing in synchrony with an auditory signal

Waka Fujisaki^{1,‡}, Ansgar Koene², Derek Arnold^{2,†}, Alan Johnston²
and Shin'ya Nishida^{1,*}

¹*NTT Communication Science Laboratories, NTT Corporation, Atsugi, Kanagawa 243-0198, Japan*

²*Department of Psychology, University College London, Gower Street, London WC1E 6BT, UK*

We examined whether the detection of audio–visual temporal synchrony is determined by a pre-attentive parallel process, or by an attentive serial process using a visual search paradigm. We found that detection of a visual target that changed in synchrony with an auditory stimulus was gradually impaired as the number of unsynchronized visual distractors increased (experiment 1), whereas synchrony discrimination of an attended target in a pre-cued location was unaffected by the presence of distractors (experiment 2). The effect of distractors cannot be ascribed to reduced target visibility nor can the increase in false alarm rates be predicted by a noisy parallel processing model. Reaction times for target detection increased linearly with number of distractors, with the slope being about twice as steep for target-absent trials as for target-present trials (experiment 3). Similar results were obtained regardless of whether the audio–visual stimulus consisted of visual flashes synchronized with amplitude-modulated pips, or of visual rotations synchronized with frequency-modulated up–down sweeps. All of the results indicate that audio–visual perceptual synchrony is judged by a serial process and are consistent with the suggestion that audio–visual temporal synchrony is detected by a ‘mid-level’ feature matching process.

Keywords: cross modal perception; visual search; audio–visual synchrony

1. INTRODUCTION

The world around us is filled with auditory and visual signals. Some of these arise from the same physical events, such as the visual movement of a drumstick and the sound of a beat, while many other auditory and visual signals are unrelated. Physical synchrony, or simultaneity, is an important cue to causal relationships, generally implying a common cause. Previous studies have revealed several intriguing properties of audio–visual synchrony perception (e.g. Shipley 1964; Dixon & Spitz 1980; Lewkowicz 1996; Munhall *et al.* 1996; Sekuler *et al.* 1997; Stone *et al.* 2001; Recanzone 2003; Sugita & Suzuki 2003; Alais & Burr 2004; Fujisaki *et al.* 2004; Kopinska & Harris 2004; Lewald & Guski 2004; Vroomen *et al.* 2004; Arnold *et al.* 2005; Fujisaki & Nishida 2005), but most studies have examined conditions in which only a small number of audio–visual events have existed at any moment in time (such as lip movement and speech, a flash and a pip, or a bouncing ball and a collision sound). It is unclear how we find synchronous audio–visual pairs in temporally cluttered audio–visual environments while avoiding false audio–visual matches and combinational explosions.

To address this issue, we conducted a series of visual search experiments in which several dynamic visual stimuli were simultaneously presented in a display. The changes imposed on each visual stimulus were not correlated. The participant had to detect which visual

target changed in synchrony with an auditory stimulus. Our main interest was the set-size effect—how search performance changes with number of uncorrelated visual stimuli (distractors). If search performance were independent of distractor number a parallel search would be implicated in which one can simultaneously check many audio–visual signals for possible pairings. This would also suggest that, when present, a synchronous target should rapidly become apparent and may appear to ‘pop out’ from the distractors. On the other hand, if search speed is positively correlated with distractor number, a ‘serial search’ would be implicated in which participants can only check one or a small number of signals at a time for possible pairings.

A ‘parallel search’ is expected if audio–visual synchrony detection is a pre-attentive process that consumes few attentional resources. If this were the case it is likely that audio–visual synchrony would be detected by specialized mechanisms at an early stage of sensory processing. There are a few lines of evidence that are consistent with this possibility. For instance, the ventriloquist effect, an illusory visual capture of the spatial location of an auditory signal, may occur pre-attentively since the illusion helps intra-modal segmentation of two auditory streams presented from the same location, by making them appear to come from different locations (Driver 1996). The ventriloquist effect can also be induced by unattended visual stimuli when voluntary (Bertelson *et al.* 2000) or stimulus-driven (Vroomen *et al.* 2001) attention is directed to another location. Additionally, auditory attention can be drawn to the location of a visual cue when it is paired with a concurrent unlocalizable sound, to produce ventriloquism (Spence & Driver 2000). Other

* Author for correspondence (nishida@brl.ntt.co.jp).

† Present address: Department of Psychology, University of Sydney, Sydney, NSW 2006, Australia.

‡ W.F. is a Research Fellow of the Japan Society for the Promotion of Science.

phenomena that could be interpreted as suggesting early pre-attentive binding of audio-visual signals include the enhanced audibility/visibility of coupled audio-visual signals (Stein *et al.* 1996; Odgaard *et al.* 2004; Sheth & Shimojo 2004; but also see Odgaard *et al.* 2003), perceptual integration of visual and auditory motion signals (Meyer *et al.* 2005; Soto-Faraco *et al.* 2005; but also see Wuerger *et al.* 2003; Alais & Burr 2004), visual modulation of auditory perception (McGurk & MacDonald 1976; Soto-Faraco *et al.* 2004) and auditory modulation of visual perception (Sekuler *et al.* 1997; Shimojo & Shams 2001).

On the other hand, a 'serial search' would be expected if detection of audio-visual synchrony were determined by an attentive process. This possibility is consistent with the hypothesis that audio-visual synchrony is detected by a mid-level general-purpose mechanism (Fujisaki & Nishida 2005), as well as with the classical theory that the integration of separately analysed features requires attention (Treisman & Gelade 1980).

In the first experiment, we measured the accuracy of detecting an audio-visually correlated target. The results showed a clear set-size effect—as the number of visual elements increased, target detection performance decreased. The set-size effect was caused by limited attentional resources and not perceptual interference since audio-visual synchrony detection accuracy was not influenced by distractor number when a 'to be attended' position was pre-cued. The third experiment showed a linear increase of response time to target detection as the number of distractors increased. All of the results indicate that audio-visual synchrony is determined by a serial process.

2. EXPERIMENT 1

The first experiment measured the accuracy of detecting a visual target whose modulation was temporally synchronized with the modulation of an auditory signal. Stimulus exposure duration was fixed. To confirm the generality of the finding, we used two types of audio-visual stimuli. One consisted of visual flashes synchronized with amplitude-modulated (AM) pips (figure 1), the other comprised visual rotations synchronized with frequency-modulated (FM) sweeps (figure 2).

(a) Method

(i) Participants

Participants were two of the authors and four paid volunteers who were unaware of the purpose of the experiments. One of the volunteers participated in the flash-pip condition only and another participated in the rotation-sweep condition only. All had normal or corrected-to-normal vision and hearing. Informed consent was obtained before the experiment started.

(ii) Apparatus

The apparatus was identical to that used in Fujisaki & Nishida (2005). In brief, visual stimuli were generated with a VSG2/5 (Cambridge Research Systems), and auditory stimuli were generated with a TDT Basic Psychoacoustic Workstation (Tucker-Davis Technologies). The participant sat in a quiet dark room 57 cm from a

monitor (SONY GDM-F500, frame rate: 160 Hz) while wearing headphones (Sennheiser HDA 200).

(iii) Stimuli

A visual display for the flash-pip condition consisted of 2, 4 or 8 luminance-modulated Gaussian blobs (s.d. = 0.5°) arranged horizontally (when $N=2$), in a square ($N=4$), or in a circle ($N=8$) at an eccentricity of 7.53° from a central bull's-eye fixation point (see figure 1a). The background was a 21.5 cd m^{-2} uniform field subtending 38.7° in width and 29.5° in height. Each blob appeared only when the pulse was 'on'. The luminance increment at the centre relative to the background was 43 cd m^{-2} . Gaussian blobs consist predominantly of low-spatial frequencies, to which visual responses are rapid (Kelly 1979). The auditory stimulus was white noise presented diotically via headphones with a sampling frequency of 24 420 Hz. The noise was presented with intensity $\sim 54 \text{ dB SPL}$ and was only delivered when the pulse was 'on'. During a stimulus presentation lasting 2 s, the amplitudes of the auditory and visual stimuli were modulated by random pulse trains. The pulse trains that modulated different visual elements were uncorrelated (figure 1b). The modulation of the auditory signal was identical to the modulations of one of the visual stimuli—the target (see figure 1c). Generation of a random pulse train was based on the refresh rate of the monitor (160 Hz). For every frame (6.25 ms) the pulse was 'on', independent of the state of previous frames, with a probability of 3.13, 6.25, 12.50 or 25%, which respectively resulted in pulse densities of 5, 10, 20 and 40 pulses s^{-1} . Pulse modulation of the auditory stimulus was made by upsampling the 160 Hz target visual pulse to 24 420 Hz, and setting a pulse duration of 6.25 ms.

The visual display for the rotation-sweep stimulus consisted of 2, 4 or 8 windmills (3.72° in diameter; figure 2a). Each windmill had a radial sinusoidal luminance modulation of 4 cycles per rotation, with 50% luminance modulation from the background (43 cd m^{-2}). During a stimulus presentation lasting 2 s, each windmill rotated at a constant speed (180° s^{-1} in terms of rotation speed) and reversed at a rate of 10 times s^{-1} . Reversal timings were randomly determined with the constraints that rotation was confined to $\pm 90^\circ$ from the initial orientation and that the total moving time was equal for clockwise and anticlockwise rotations. The auditory stimulus was a pure tone. The frequency was logarithmically increased or decreased from 440 Hz at a speed of 2 octaves s^{-1} . The sweep direction was reversed 10 times s^{-1} with the constraints that frequency should not exceed the range between 220 and 880 Hz (± 1 octave from 440 Hz), and returned to 440 Hz at the end of a 2 s presentation. Sound level was about 54 dB SPL when frequency was at 440 Hz. Since the headphones we used are known to have relatively flat frequency characteristics (cf. Hirahara 2004), we did not compensate for sound level across frequencies. Reversal timings were uncorrelated among different windmills. The trajectory of the FM sweep was synchronized with the movements of one of the visual stimuli (figure 2c). Clockwise and anticlockwise windmill rotations corresponded to upward and downward FM sweeps of the auditory stimulus, respectively. The initial orientation of each windmill was randomized to prevent participants from making judgments based on

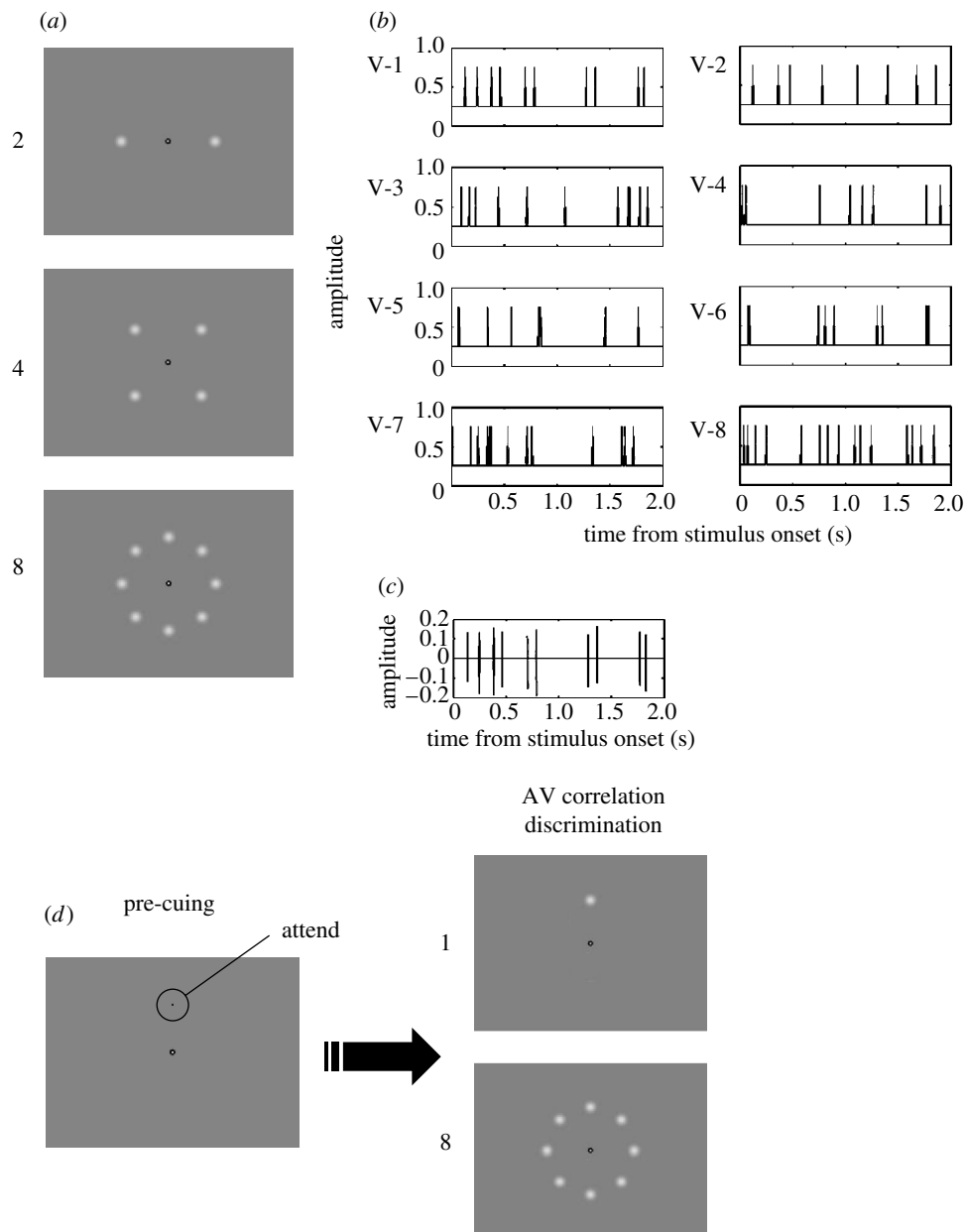


Figure 1. Flash-pip stimuli used in experiments 1 and 2. (a) Visual displays used in experiment 1 consisted of 2, 4 or 8 luminance-modulated Gaussian blobs (s.d. = 0.5°) arranged horizontally (when $N=2$), in a square ($N=4$), or circularly ($N=8$) at an eccentricity of 7.53° from a central fixation point. (b) Example of luminance-modulated waveforms used for the eight visual elements condition. In this example the visual stream labelled V-1 is temporally correlated with the auditory stimulus. (c) An example of the auditory stimulus (AM white noise). (d) Visual displays used in experiment 2. The number of visual stimuli was either 1 (a target only) or 8 (a target and seven distractors). When eight stimuli were shown, target location was randomly chosen from eight possible locations and was pre-cued by a probe (a black circle 0.30° in diameter).

the relationship between the absolute rotation angle and absolute pitches of the auditory stimulus.

(iv) Procedure

For each trial, an audio-visual sequence was presented for 2 s. The sound-synchronized visual target was presented at a random position. The participant had to indicate the target position by clicking one of the buttons displayed on a second monitor screen. The spatial layout of the response buttons was the same as that of the visual stimuli. Feedback was provided, after participants had made a response, by showing a square black frame ($0.4 \times 0.4^\circ$) at the target position for 2.2 s (flash-pip) or for approximately 5 s (rotation-sweep). The inter-trial interval was 0.3 s, during which only the fixation point was shown.

There were two within-participants factors: the number of visual elements (3 levels: 2, 4 and 8) and stimulus density (4 levels: 5, 10, 20 and 40 pulses s^{-1} for the flash-pip condition, 1 level: $10 \text{ reversal s}^{-1}$ for the rotation-sweep condition). These parameters were fixed within each block. Each participant completed at least two blocks for each stimulus condition. One block consisted of 20 trials, plus four initial practice trials.

(b) Results and discussion

Figure 3a shows the proportions correct for target detection of each condition. This value, however, cannot be directly compared between conditions since chance performance decreases with set size. We therefore computed d' for N alternative forced choice (Hacker & Ratcliff

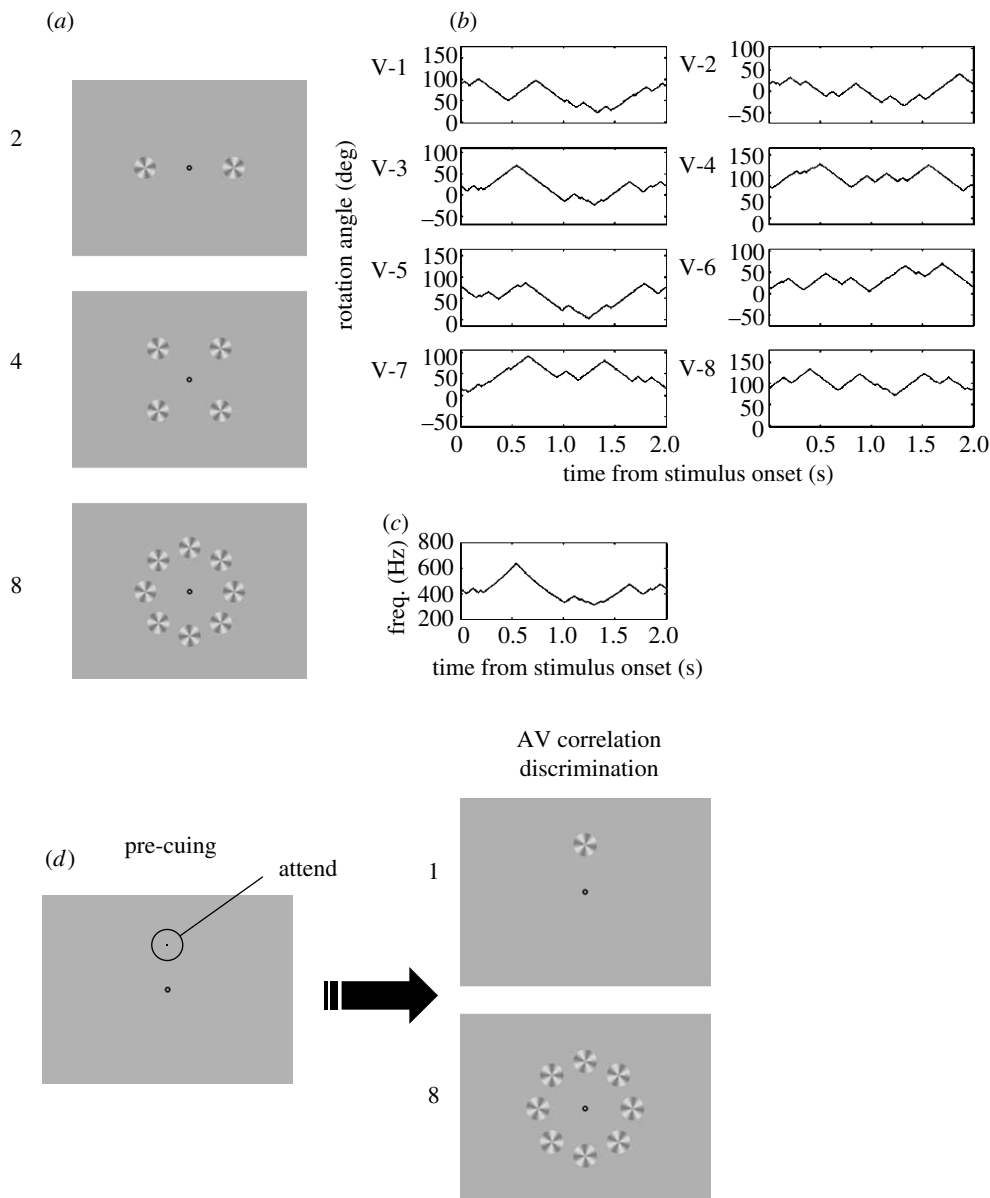


Figure 2. Rotation-sweep stimuli used in experiments 1 and 2. (a) Visual displays used in experiment 1 consisted of 2, 4 or 8 windmills (3.72° in diameter) arranged horizontally ($N=2$), in a square ($N=4$), or circularly ($N=8$) at an eccentricity of 7.53° from the fixation point. (b) An example of modulated waveforms used for the eight visual elements condition. V-3 is temporally correlated with the auditory stimulus. (c) An example of the auditory stimuli (FM sweep, logarithmically increasing). (d) Visual displays used in experiment 2. The location of the target was pre-cued by a probe.

1979)¹. For chance-level performance, $d' = 0$ regardless of the set size (N).

Figure 3c (solid lines) shows d' for flash-pip stimuli (separately for four different densities) averaged over the five participants, as a function of the number of the visual elements. A two-way analysis of variance indicated that the main effects of visual element number [$F(2,8) = 19.04$, $p < 0.01$] and temporal density [$F(3,12) = 62.19$, $p < 0.01$], and their interaction [$F(6,24) = 9.94$, $p < 0.01$], were all significant. Target detection was impaired as pulse temporal density increased. A more detailed study on the effects of pulse density will be reported elsewhere. The main point here is that there was a consistent trend across difficulty level that target detection performance became poorer as the number of distracters increased. Set-size effects were significant [5 s^{-1} : $F(2,8) = 75.02$, $p < 0.01$, 10 s^{-1} : $F(2,8) = 21.84$, $p < 0.01$, 20 s^{-1} : $F(2,8) = 5.02$, $p = 0.039$], except for the highest density condition [40 s^{-1} : $F(2,8) = 2.13$, $p > 0.10$] where performance was

close to chance. These results suggest that audio-visual synchrony search is not 'parallel', but 'serial'. An audio-visual synchronous target does not 'pop-out'.

One concern about the flash-pip stimulus is that Gaussian blobs flashing with appropriate timing could be spatiotemporally grouped, thereby eliciting apparent motion. There are some reports showing that within-modal grouping can suppress crossmodal binding (Watanabe & Shimojo 2001; Sanabria *et al.* 2005). It is known that visual motion perception is affected by auditory stimuli (Sekuler *et al.* 1997; Meyer & Wuerger 2001; Watanabe & Shimojo 2001) and vice versa (Soto-Faraco *et al.* 2002; Sanabria *et al.* 2005). In addition, flash-pip combinations can elicit audio-visual illusions, including the double flash illusion (Shams *et al.* 2000) and temporal ventriloquism (Morein-Zamir *et al.* 2003), in which apparent visual temporal patterns are captured by auditory stimuli. These factors might make visual searches particularly difficult for flash and pip

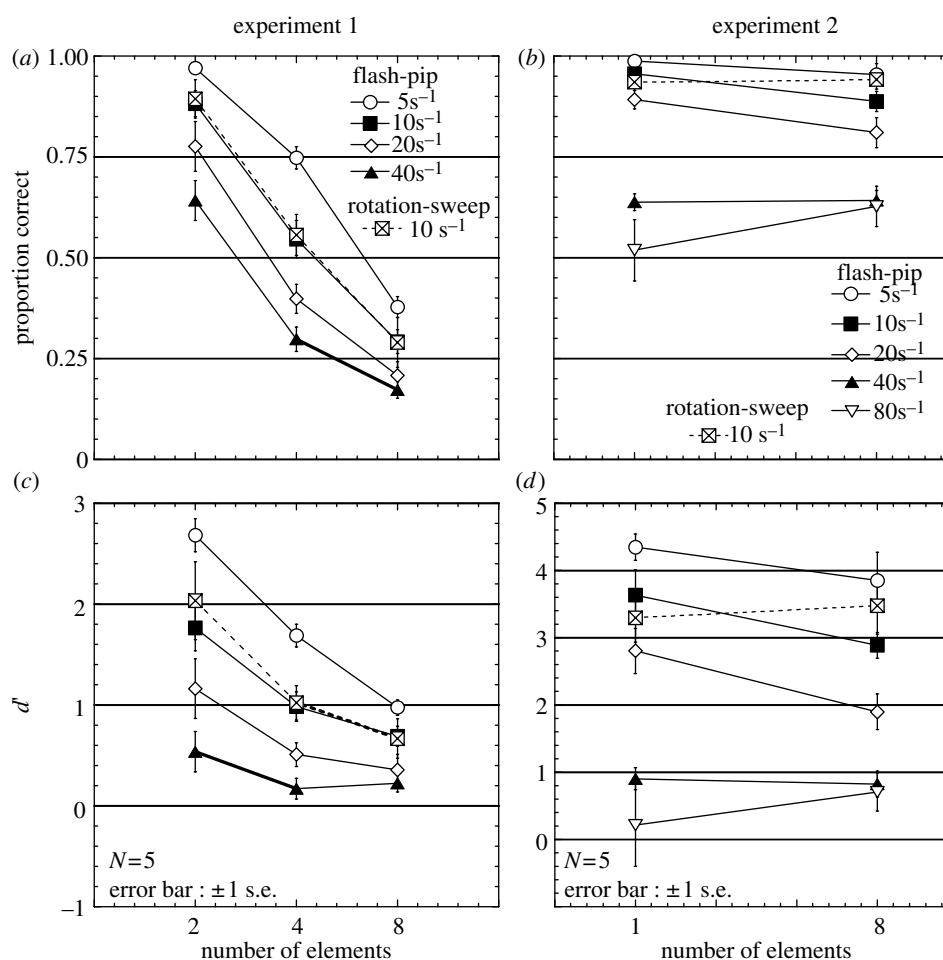


Figure 3. (a) Results obtained in experiment 1. The proportion correct for different densities of flash-pip stimuli (5, 10, 20, 40 pulses s^{-1}) and rotation-sweep stimuli (10 reversals s^{-1}) as a function of the number of visual elements (2, 4, 8). (c) The same data expressed in terms of d' . (b) Proportion correct obtained in experiment 2 (synchrony discrimination of pre-cued target) as a function of the number of visual elements (1, 8). (d) The same data expressed in terms of d' .

stimuli. However, we obtained a very similar set-size effect even when we paired windmill rotations with FM sweeps (figure 3*a,c*, dotted line; effect of number of visual elements: $F(2,8)=12.64$, $p<0.01$), although the FM sweep reversals had no transient pulse that might induce a phenomenon like illusory double flashes, and rotation reversals did not induce apparent motion between windmills.

3. EXPERIMENT 2

The set-size effect shown in experiment 1 could be interpreted as indicating a serial search for audio-visual synchrony that consumes significant amounts of attentional resources. However, distractors could impair target detection in a different way by perceptually reducing the visibility of the target through lateral masking or by a crowding effect (Loomis 1978). In experiment 2, we presented a cue to target location prior to stimulus presentations and asked participants to judge whether that target was synchronous with the auditory stimulus. If set-size effects reflect an attentional limitation synchrony judgments should not be impaired by the presence of distractors when the participant attends to the target. If set-size effects reflect perceptual interference, the presence of distractors should still impair target synchrony discrimination.

(a) Method

Methods in experiment 2 were the same as those used in experiment 1 with the following exceptions. Either one (a target only) or eight (a target and seven distractors) visual stimuli were presented. In the latter condition, the target location, which was randomly chosen from eight possible locations, was pre-cued by a black circle (0.30° in diameter) presented for 2.2 s (flash-pip) or for approximately 5 s (rotation-sweep) until 0.3 s before the stimulus presentation (see figures 1*d* and 2*d*). The modulation of the pre-cued visual target was either synchronized (in a half of the trials) or uncorrelated (in the other half) with the modulation of the auditory stimulus. Distractor modulations were always uncorrelated with the sound modulation. Participants indicated whether the target was synchronous with the sound by pressing one of two buttons on a second monitor screen. Feedback was given after each response by colour changes of the fixation marker; blue indicated 'synchronized', red 'uncorrelated'.

(b) Results and discussion

Figure 3*b,d* show the discrimination performance in terms of proportion correct and d' for flash-pip stimuli (separately for five different densities) and a rotation-sweep stimulus (10 reversals s^{-1}), averaged over the five participants and plotted as a function of the number of visual elements. Two-way analyses of variance of data

from the flash-pip condition showed that the main effect of density was significant for both proportion correct ($F(4,16)=35.01$, $p<0.01$) and d' ($F(4,16)=31.32$, $p<0.01$); the effect of visual element number was insignificant for proportion correct ($F(1,4)=1.18$, $p>0.1$), while significant for d' ($F(1,4)=11.61$, $p=0.027$); the interaction between density and element number was significant for proportion correct ($F(4,16)=4.61$, $p=0.011$), while insignificant for d' ($F(4,16)=2.33$, $p>0.10$). Although these mixed results do not completely reject a possible effect of element number for the flash-pip condition, one-way analyses of variance of data from the rotation-sweep condition (under which possible within-modal interactions were minimized) indicated no significant effect of number of visual elements for both proportion correct ($F(1,4)=0.06$, $p>0.10$) and d' ($F(1,4)=0.13$, $p>0.10$). The lack of significant differences between the one and eight element conditions shows that when participants attended to the visual target, distractors had little influence on audio-visual synchrony judgments. This suggests that the set-size effect found in experiment 1 should be ascribed to the requirement to distribute attentional resources between distractors when making synchrony judgments rather than to some distractor-induced perceptual interference.

4. UNLIMITED-CAPACITY, PARALLEL PROCESSING?

The decrease in target detection accuracy with element number (set-size effect) does not necessarily imply attentive (limited-capacity) processing. According to signal detection theory, the set-size effect could also result from noisy, low-threshold, unlimited-capacity, parallel processing (e.g. Palmer *et al.* 2000). Given that a distractor is judged to be a target (false alarm) with a given probability, the error rate should rise as the number of distractors increases. To see to what extent a pre-attentive, parallel processing model can account for our results, we compared the search performance obtained in experiment 1 with that predicted by a parallel processing model based on the hit and false alarm rates for a single audio-visual synchrony event obtained in experiment 2.

We assume that for a given presentation in the visual search experiment, each target/distractor is independently judged either as correlated (C) or uncorrelated (UC). Counting all the possible combinations of C-UC judgments, there are 2^N cases (4, 16 and 256 cases for $N=2, 4, 8$, respectively). The probability of the occurrence of each case can be obtained by multiplying the probabilities of C or UC judgment for each element. For instance, when $N=4$, the probability that the target and the 2nd distractor are judged as C and the others are judged as UC is

$$p(\text{C, UC, C, UC}) = p(\text{target} = \text{C})p(\text{distractor}_1 = \text{UC}) \\ \times p(\text{distractor}_2 = \text{C})p(\text{distractor}_3 = \text{UC}).$$

This probability is estimated from the performance for single element judgments in experiment 2: $p(\text{target} = \text{C}) = p(\text{hit})$; $p(\text{target} = \text{UC}) = p(\text{miss})$; $p(\text{distractor} = \text{C}) = p(\text{FA: false alarm})$; $p(\text{distractor} = \text{UC}) = p(\text{CR: correct rejection})$. In general, if we number each of the possible

2^N cases, the probability of occurrence of the i th cases is

$$p(i) = p(\text{target} = \text{C})^\alpha p(\text{target} = \text{UC})^{1-\alpha} \\ \times p(\text{distractor} = \text{C})^\beta p(\text{distractor} = \text{UC})^{N-1-\beta},$$

where $\alpha=1$ if, for the i th case, the target is judged C, and 0 otherwise, and β is the number of distractors that are judged C for the i th case. Assuming that the subject chooses the target among apparent C elements with equal probability, the expected probability of correctly finding the target in the i th case is

$$\hat{p}_{\text{correct}}(i) = \frac{\alpha}{\alpha + \beta}.$$

For instance, this probability is 0.5 for the case where the target and one of the distractors are judged C ($\alpha=1$, $\beta=1$). The exception is when none of the elements is judged to be C. In this case, we assume that the target is chosen from apparent UC elements with equal probability.

$$\hat{p}_{\text{correct}}(i) = \frac{1}{N}.$$

Finally, the expected proportion correct is

$$\hat{p}_{\text{correct}} = \sum_{i=1}^{2^N} p(i)\hat{p}_{\text{correct}}(i).$$

Table 1 shows proportion correct and d' (in parentheses) obtained with experiment 1 (data) and the estimated values from a parallel processing model (model) for different set-size conditions (2, 4, 8, respectively). The paired t -tests show that search performance for $N \geq 4$ was significantly worse for the data than for the model prediction, suggesting that the set-size effect found in experiment 1 was too large to be accounted for by the specified parallel, pre-attentive processing model.

Experiments 1 and 2 presented a stimulus for a fixed duration (2 s) and measured the participant's performance in terms of proportion correct. The results suggest a 'serial search' for audio-visual temporal synchrony. According to this perspective, reaction time (RT) following detection of a visual target defined by audio-visual synchrony should increase linearly with set size. Additionally, the slope of the increase in RT should be twice as steep for target-absent trials (where all visual elements have to be checked) than for target-present trials (where the target is found, on average, after half of the elements have been checked). We assess these possibilities in experiment 3.

5. EXPERIMENT 3

(a) Method

There were two within-participants factors: presence/absence of the audio-visually synchronized target and the number of visual elements (1, 2, 4, 6 and 8).

Participants and stimuli were identical to those used for the rotation-sweep condition of experiments 1 and 2, with the following exceptions. A stimulus sequence made of windmills and an FM sweep tone lasted for a maximum of 30 s. Since it took too long to generate a 30 s sequence between trials, the same 10 s sequence was seamlessly repeated three times. Visual elements for $N \leq 6$ were randomly positioned at N of the eight slots defined by the stimulus $N=8$.

In each trial a stimulus was presented until participants pressed the left button on a VSG CT3 response box either

Table 1. Comparison of data and prediction by unlimited-capacity, parallel processing model. (The proportion correct and d' value (in parentheses) obtained in experiment 1 (data), estimated values from a parallel processing model (model) and the results of paired t -test. ** $p < 0.01$, * $p < 0.05$, † $p < 0.10$.)

	$n = 2$			$n = 4$			$n = 8$		
	data	model	paired t value ($t(4)$)	data	model	paired t value ($t(4)$)	data	model	paired t value ($t(4)$)
	flash-pip	5 s^{-1} (0.970 (2.684)) 0.880 (1.764)) 0.775 (1.162)) 0.644 (0.538)) 0.894 (0.386))	0.988 (3.050) 0.956 (2.572)) 0.892 (1.838)) 0.638 (0.502)) 0.935 (0.259))	2.72† (2.88*) 1.680 (1.780)) 1.950 (1.910)) 0.190 (0.260)) 0.980 (0.670))	0.749 (1.688) 0.545 (0.984)) 0.398 (0.508)) 0.298 (0.170)) 0.556 (0.171))	0.964 (3.304) 0.904 (2.712)) 0.761 (1.820)) 0.364 (0.402)) 0.857 (0.361))	8.20** (6.19***) 4.63** (3.71*) 6.74** (4.70***) 3.58* (3.64*) 4.63** (3.20*)	0.376 (0.974) 0.291 (0.690)) 0.208 (0.356)) 0.169 (0.224)) 0.290 (0.666))	0.924 (3.454) 0.827 (2.764)) 0.601 (1.770)) 0.189 (0.320)) 0.747 (2.362))
rotation-sweep	10 s^{-1}								

when they found a target or when they judged that there was no target. Participants were asked to respond as accurately and as rapidly as possible. Reaction times (RTs) were given by the time of the button press after the onset of the stimulus presentation. Stimulus presentation stopped immediately after the participants' button press. Participants were then asked to make an unspeeded response indicating the position of the target by clicking one of the buttons displayed on a second monitor screen. Positions of the response buttons corresponded to the positions of visual stimuli. When participants judged that there was no target they pressed a 'no target' button. Feedback was given by showing a black square ($0.4 \times 0.4^\circ$) at the target position and/or by changing the colour of the fixation stimulus—blue for 'target present' and red for 'target absent'. When participants made no response for 30 s, the response was regarded as 'target absent' with a RT of 30 s. Inter-trial intervals were approximately 16 s.

One session consisted of 20 trials, two trials for each combination of the two factors. Each participant completed 10 sessions containing 200 individual trials.

(b) Results and discussion

Figure 4 shows RTs (figure 4a,b) and proportions correct (figure 4c,d) for target present (figure 4a,c) and target absent (figure 4b,d) trials as a function of visual element number. A two-way analysis of variance was conducted on the RT data. Since the interaction between set sizes and the existence of a target was significant [$F(4,16) = 47.34$, $p < 0.01$], the main effects for each factor were analysed separately. For both target present and absent conditions, the effect of set size was significant [target present, $F(4,16) = 30.93$, $p < 0.01$; target absent, $F(4,16) = 113.68$, $p < 0.01$]. For all set-size conditions, the effect of target present/absent was also significant ($N = 1$, $F(1,4) = 9.37$, $p < 0.05$; $N = 2$, $F(1,4) = 15.18$, $p < 0.05$; $N = 4$, $F(1,4) = 17.00$, $p < 0.05$; $N = 6$, $F(1,4) = 670.17$, $p < 0.01$; $N = 8$, $F(1,4) = 216.23$, $p < 0.01$). Table 2 shows the values of the slope, intercept and R^2 of linear regressions fitted to data obtained for each subject for both target present and target absent trials. The results indicate a clear set-size effect that can be fitted well using linear regression. The slope was approximately twice as steep for the target absent condition relative to the target present condition. Proportion correct was generally high, falling off slightly at larger set sizes. This implies that the increase in RT for larger set sizes cannot be ascribed to a speed-accuracy trade off.

These results strongly support the hypothesis that search for a visual target defined by audio synchrony is a serial process. We also obtained a similar result using a stimulus made of visual flashes and auditory pips (Koene *et al.* 2005). One may point out that even for an ideal observer who can use all of the information contained by the stimulus, the latency to detect a target would become longer as the number of distractors increases. This is because the target cannot be uniquely identified until all distractors behave differently from the target. However, considering the temporal properties of our stimulus (10 changes s^{-1} occurring randomly), it would not take more than a few hundred milliseconds to uniquely decide on a target even under the condition of $N = 8$. In addition, the unlimited capacity parallel processing model cannot predict the RT data either, since it predicts that the subject should correctly choose a synchrony target within 2 s in

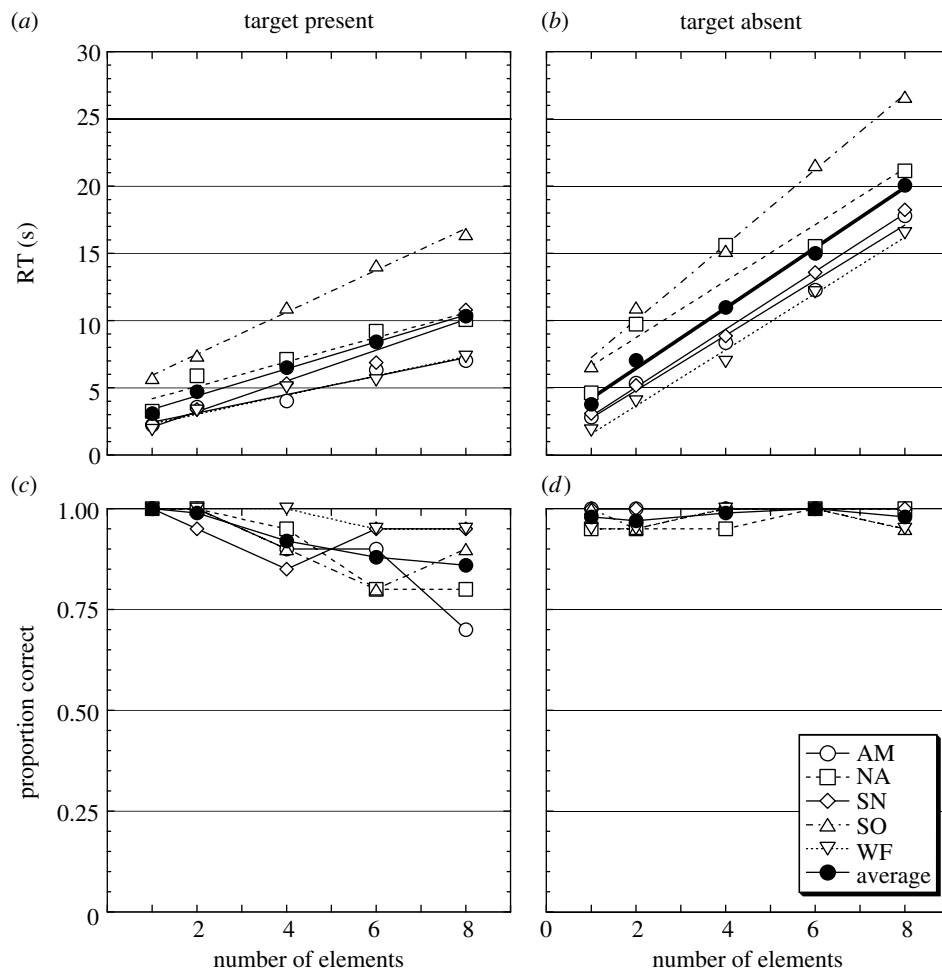


Figure 4. Reaction times (*a, b*) and proportion correct (*c, d*) for target present (*a, c*) and target absent (*b, d*) conditions as a function of number of visual elements. Lines in (*a*) and (*b*) are linear regressions obtained for each participant and for the grouped average. Open symbols show data points obtained with individual participants. Filled symbols show the grouped averages.

Table 2. The values of the slope, intercept and R^2 of linear regressions fitted to data for each subject for target present and absent conditions in experiment 3. (The group mean represents the linear fit for averaged grouped data. Slope ratios indicate the ratio of slopes fitted to target absent and target present conditions.)

subject	target present			target absent			slope ratio
	slope	intercept	R^2	slope	intercept	R^2	
AM	0.682	1.772	0.956	2.050	0.713	0.988	3.008
NA	0.914	3.270	0.932	2.103	4.497	0.907	2.301
SN	1.141	0.954	0.966	2.154	0.735	0.998	1.888
SO	1.558	4.395	0.994	2.800	4.442	0.994	1.798
WF	0.715	1.601	0.960	2.091	-0.520	0.992	2.923
group mean	1.002	2.398	0.994	2.240	1.973	0.996	2.236

74.7% of the trials even when $N=8$ (table 1). This performance is obviously better than the obtained RTs, which are closer to 10 s.

6. GENERAL DISCUSSION

This study examined whether audio-visual temporal synchronies for multiple objects are pre-attentively processed in parallel or attentively processed in serial. We found that the detection of a visual target changing in synchrony with an auditory stimulus was gradually impaired as the number of unsynchronized visual distractors increased (experiment 1), whereas synchrony

discrimination of an attended target was unaffected by the presence of distractors (experiment 2). The effects of distractors cannot be ascribed to reduced visibility of the target, nor can the increase in false alarm rates be predicted by a noisy parallel processing model. RTs to find a target increased linearly with number of distractors (experiment 3). Similar results were obtained regardless of whether the audio-visual stimulus was visual flashes synchronized with AM pips, or visual rotations synchronized with FM sweeps. All of the present results suggest that audio-visual temporal synchrony detection is a serial process.²

In our experiments, the target visual stimulus was synchronized with the auditory stimulus with no physical

delay. Considering various factors that could affect the point of audio-visual simultaneity (Stone *et al.* 2001; Fujisaki *et al.* 2004), whether physical simultaneity is the optimal condition for discriminating correlated stimuli from uncorrelated ones is an interesting open question. This issue, however, does not require a caveat to our conclusion given that our subjects found no difficulty in discriminating physically synchronous stimuli from asynchronous ones when they knew the location of the target (experiment 2).

The present results, together with a previous finding that distinguishing between repetitive auditory and visual signals alternating in synchrony, or with timing differences, is greatly impaired when the alternations are rapid (Fujisaki & Nishida 2005), suggest that audio-visual temporal synchrony is not detected by early low-level specialized sensors. On the other hand, synchrony is unlikely to be detected by a highly cognitive process given that subjective audio-visual synchrony can be recalibrated by the constant time lag between visual and auditory delays (Fujisaki *et al.* 2004; Vroomen *et al.* 2004).³ Rather, we hypothesize that audio-visual synchrony is detected by a general-purpose mid-level perceptual mechanism that compares salient features extracted from each signal stream. In our visual search task, visual elements are equally salient, so the participant has to attentively select one after another to compare it with the auditory signal. The temporal limitation of audio-visual synchrony detection (Fujisaki & Nishida 2005) can be ascribed to a difficulty in individuating salient features within rapid stimulus sequences. Our hypothesis is consistent with prior findings—that within-modal grouping precedes cross-modal binding (Watanabe & Shimojo 2001; Sanabria *et al.* 2005) and that audio-visual speech integration (which had been considered to occur pre-attentively) is greatly impaired under high attentional load (Alsius *et al.* 2005). Additionally, our hypothesis implies homology of processing styles (i.e. matching salient features) for cross-modal binding and for within-modal (visual) cross-attribute binding (Treisman & Gelade 1980; Holcombe & Cavanagh 2001; Lu & Sperling 2001; Nishida & Johnston 2002; Arrighi *et al.* 2005). One might suggest that we could not find low-level audio-visual mechanisms since our audio-visual signals are not spatially co-localized (Meyer *et al.* 2005) or they are not sufficiently close to real events (e.g. looming, speech). Although we cannot exclude these possibilities, as far as we know, there is no evidence that spatial co-localization or natural binding dramatically improve audio-visual synchrony perception. Note also that it would be technically difficult to test the effect of co-localization using the standard visual search paradigm.

According to our hypothesis, audio-visual binding is not always attention demanding. Under environments containing a small number of events, early 'bottom-up' segmentation processes for each modality can unambiguously extract corresponding audio and visual signals as salient features. Hypothetically, in these circumstances the mid-level perceptual process can detect audio-visual synchrony without consuming attentional resources. Previous studies showing 'pre-attentive' audio-visual binding have generally used this type of stimulus configuration (Bertelson *et al.* 2000; Spence & Driver 2000; Vroomen *et al.* 2001). Once an audio-visual binding has been established virtually 'automatically' at the

onset of the stimulus, it might be resistant to modulation by attention (Bertelson *et al.* 2000; Vroomen *et al.* 2001). Our hypothesis does not exclude the possibility that the results of cross-modal binding may affect subsequent attentional processes (Spence & Driver 2000). Cross-modal binding may also modulate within-modal segregation (Driver 1996), possibly through feedback mechanisms which might also contribute to other audio-visual interactions that are often cited as evidence for modulation of 'early' or low level perceptual processes (McGurk & MacDonald 1976; Stein *et al.* 1996; Shimojo & Shams 2001; Sheth & Shimojo 2004; Soto-Faraco *et al.* 2004; Meyer *et al.* 2005; Soto-Faraco *et al.* 2005). However, it is entirely possible that some of these phenomena might also be mediated at the same processing stage as, or even later than, the saliency-based cross-modal binding process that we argue for here.

This work was supported by NTT and the Human Frontier Science Program.

ENDNOTES

¹In transforming proportion correct (p) to d' , we replaced $p=1.0$ with $p=0.99$ to avoid an infinite d' .

²Independent of our study, van de Par & Kohlrausch (2004) have conducted a similar visual search experiment using moving discs and an AM tone. They reported that RTs to find a visually synchronized auditory target increased with the number of visual distractors. In addition, they found a similar result when the roles of visual and auditory stimuli were reversed. That is, the RT to search for a visually synchronized auditory target presented with a given pitch increased with the number of auditory distractors presented at different frequencies. The findings of van de Par & Kohlrausch (2004) therefore support our conclusion of serial processing of audio-visual temporal synchrony.

³Recalibration of simultaneity by the distance of the audio-visual event from the observer is also reported (Engel & Dougherty 1971; Sugita & Suzuki 2003; Kopinska & Harris 2004; Alais & Carlile 2005; but see also Stone *et al.* 2001; Lewald & Guski 2004; Arnold *et al.* 2005).

REFERENCES

- Alais, D. & Burr, D. 2004 No direction-specific bimodal facilitation for audio-visual motion detection. *Brain Res. Cogn. Brain Res.* **19**, 185–194. (doi:10.1016/j.cogbrainres.2003.11.011)
- Alais, D. & Carlile, S. 2005 Synchronizing to real events: subjective audio-visual alignment scales with perceived auditory depth and speed of sound. *Proc. Natl Acad. Sci. USA* **102**, 2244–2247.
- Alsius, A., Navarra, J., Campbell, R. & Soto-Faraco, S. 2005 Audiovisual integration of speech falters under high attention demands. *Curr. Biol.* **15**, 839–843. (doi:10.1016/j.cub.2005.03.046)
- Arnold, D. H., Johnston, A. & Nishida, S. 2005 Timing sight and sound. *Vision Res.* **45**, 1275–1284. (doi:10.1016/j.visres.2004.11.014)
- Arrighi, R., Alais, D. & Burr, D. 2005 Perceived timing of first and second order changes in vision and audition. *Exp. Brain Res.* **166**, 445–454.
- Bertelson, P., Vroomen, J., de Gelder, B. & Driver, J. 2000 The ventriloquist effect does not depend on the direction of deliberate visual attention. *Percept. Psychophys.* **62**, 321–332.
- Dixon, N. F. & Spitz, L. 1980 The detection of auditory visual desynchrony. *Perception* **9**, 719–721.
- Driver, J. 1996 Enhancement of selective listening by illusory mislocation of speech sounds due to lip-reading. *Nature* **381**, 66–68. (doi:10.1038/381066a0)
- Engel, G. R. & Dougherty, W. G. 1971 Visual-auditory distance constancy. *Nature* **234**, 308.

- Fujisaki, W. & Nishida, S. 2005 Temporal frequency characteristics of synchrony–asynchrony discrimination of audio–visual signals. *Exp. Brain Res.* **166**, 455–464.
- Fujisaki, W., Shimojo, S., Kashino, M. & Nishida, S. 2004 Recalibration of audio–visual simultaneity. *Nat. Neurosci.* **7**, 773–778. (doi:10.1038/nm1268)
- Hacker, M. & Ratcliff, R. 1979 A revised table of d' for M-alternative forced choice. *Percept. Psychophys.* **26**, 168–170.
- Hirahara, T. 2004 Physical characteristics of headphones used in psychophysical experiments. *Acoust. Sci. & Tech.* **25**, 276–285. (doi:10.1250/ast.25.276)
- Holcombe, A. O. & Cavanagh, P. 2001 Early binding of feature pairs for visual perception. *Nat. Neurosci.* **4**, 127–128. (doi:10.1038/83945)
- Kelly, D. H. 1979 Motion and vision. II. Stabilized spatio-temporal threshold surface. *J. Opt. Soc. Am.* **69**, 1340–1349.
- Koene, A., Fujisaki, W., Arnold, D. H., Johnston, A. & Nishida, S. 2005 Cross modal correlation search in the presence of visual distractors. [Abstract.] *J. Vision* **5**, 876a. See <http://journalofvision.org/5/8/876>. (doi:10.1167/5.8.876)
- Kopinska, A. & Harris, L. R. 2004 Simultaneity constancy. *Perception* **33**, 1049–1060. (doi:10.1068/p5169)
- Lewald, J. & Guski, R. 2004 Auditory–visual temporal integration as a function of distance: no compensation for sound-transmission time in human perception. *Neurosci. Lett.* **357**, 119–122. (doi:10.1016/j.neulet.2003.12.045)
- Lewkowicz, D. J. 1996 Perception of auditory–visual temporal synchrony in human infants. *J. Exp. Psychol. Hum. Percept. Perform.* **22**, 1094–1106. (doi:10.1037/0096-1523.22.5.1094)
- Loomis, J. M. 1978 Lateral masking in foveal and eccentric vision. *Vision Res.* **18**, 335–338. (doi:10.1016/0042-6989(78)90168-2)
- Lu, Z. L. & Sperling, G. 2001 Three-systems theory of human visual motion perception: review and update. *J. Opt. Soc. Am. A Opt. Image Sci. Vis.* **18**, 2331–2370.
- McGurk, H. & MacDonald, J. 1976 Hearing lips and seeing voices. *Nature* **264**, 746–748. (doi:10.1038/264746a0)
- Meyer, G. F. & Wuerger, S. M. 2001 Cross-modal integration of auditory and visual motion signals. *Neuroreport* **12**, 2557–2560. (doi:10.1097/00001756-200108080-00053)
- Meyer, G. F., Wuerger, S. M., Rohrbein, F. & Zetzsche, C. 2005 Low-level integration of auditory and visual motion signals requires spatial co-localisation. *Exp. Brain Res.* **166**, 538–547.
- 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)
- Munhall, K. G., Gribble, P., Sacco, L. & Ward, M. 1996 Temporal constraints on the McGurk effect. *Percept. Psychophys.* **58**, 351–362.
- Nishida, S. & Johnston, A. 2002 Marker correspondence, not processing latency, determines temporal binding of visual attributes. *Curr. Biol.* **12**, 359–368. (doi:10.1016/S0960-9822(02)00698-X)
- Odgaard, E. C., Arieh, Y. & Marks, L. E. 2003 Cross-modal enhancement of perceived brightness: sensory interaction versus response bias. *Percept. Psychophys.* **65**, 123–132.
- Odgaard, E. C., Arieh, Y. & Marks, L. E. 2004 Brighter noise: sensory enhancement of perceived loudness by concurrent visual stimulation. *Cogn. Affect. Behav. Neurosci.* **4**, 127–132.
- Palmer, J., Verghese, P. & Pavel, M. 2000 The psychophysics of visual search. *Vision Res.* **40**, 1227–1268. (doi:10.1016/S0042-6989(99)00244-8)
- Recanzone, G. H. 2003 Auditory influences on visual temporal rate perception. *J. Neurophysiol.* **89**, 1078–1093.
- Sanabria, D., Soto-Faraco, S., Chan, J. & Spence, C. 2005 Intramodal perceptual grouping modulates multisensory integration: evidence from the crossmodal dynamic capture task. *Neurosci. Lett.* **377**, 59–64. (doi:10.1016/j.neulet.2004.11.069)
- Sekuler, R., Sekuler, A. B. & Lau, R. 1997 Sound alters visual motion perception. *Nature* **385**, 308. (doi:10.1038/385308a0)
- Shams, L., Kamitani, Y. & Shimojo, S. 2000 Illusions. What you see is what you hear. *Nature* **408**, 788. (doi:10.1038/35048669)
- Sheth, B. R. & Shimojo, S. 2004 Sound-aided recovery from and persistence against visual filling-in. *Vision Res.* **44**, 1907–1917. (doi:10.1016/j.visres.2004.03.009)
- Shimojo, S. & Shams, L. 2001 Sensory modalities are not separate modalities: plasticity and interactions. *Curr. Opin. Neurobiol.* **11**, 505–509. (doi:10.1016/S0959-4388(00)00241-5)
- Shiple, T. 1964 Auditory flutter-driving of visual flicker. *Science* **145**, 1328–1330.
- Soto-Faraco, S., Lyons, J., Gazzaniga, M., Spence, C. & Kingstone, A. 2002 The ventriloquist in motion: illusory capture of dynamic information across sensory modalities. *Brain Res. Cogn. Brain Res.* **14**, 139–146. (doi:10.1016/S0926-6410(02)00068-X)
- Soto-Faraco, S., Navarra, J. & Alsius, A. 2004 Assessing automaticity in audio–visual speech integration: evidence from the speeded classification task. *Cognition* **92**, B13–B23. (doi:10.1016/j.cognition.2003.10.005)
- Soto-Faraco, S., Spence, C. & Kingstone, A. 2005 Assessing automaticity in the audio–visual integration of motion. *Acta Psychol. (Amst)* **118**, 71–92. (doi:10.1016/j.actpsy.2004.10.008)
- Spence, C. & Driver, J. 2000 Attracting attention to the illusory location of a sound: reflexive crossmodal orienting and ventriloquism. *Neuroreport* **11**, 2057–2061.
- Stein, B. E., London, N., Wilkinson, L. K. & Price, D. D. 1996 Enhancement of perceived visual intensity by auditory stimuli: a psychophysical analysis. *J. Cogn. Neurosci.* **8**, 497–506.
- Stone, J. V., Hunkin, N. M., Porrill, J., Wood, R., Keeler, V., Beanland, M., Port, M. & Porter, N. R. 2001 When is now? Perception of simultaneity. *Proc. R. Soc. B* **268**, 31–38. (doi:10.1098/rspb.2000.1326)
- Sugita, Y. & Suzuki, Y. 2003 Audiovisual perception: implicit estimation of sound-arrival time. *Nature* **421**, 911. (doi:10.1038/421911a)
- Treisman, A. M. & Gelade, G. 1980 A feature-integration theory of attention. *Cognit. Psychol.* **12**, 97–136. (doi:10.1016/0010-0285(80)90005-5)
- van de Par, S. & Kohlrausch, A. 2004 Visual and auditory object selection based on temporal correlations between auditory and visual cues. In *Proc. 18th Int. Congress on Acoustics*. Kyoto, Japan, 4–9 April 2004, pp. 2055–2058.
- Vroomen, J., Bertelson, P. & de Gelder, B. 2001 The ventriloquist effect does not depend on the direction of automatic visual attention. *Percept. Psychophys.* **63**, 651–659.
- 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)
- Watanabe, K. & Shimojo, S. 2001 When sound affects vision: effects of auditory grouping on visual motion perception. *Psychol. Sci.* **12**, 109–116. (doi:10.1111/1467-9280.00319)
- Wuerger, S. M., Hofbauer, M. & Meyer, G. F. 2003 The integration of auditory and visual motion signals at threshold. *Percept. Psychophys.* **65**, 1188–1196.