Perception & Psychophysics
2007, 69 (2), 298-309

The influence of synchronous audiovisual distractors on audiovisual temporal order judgments

Argiro Vatakis and Linda Bayliss
University of Oxford, Oxford, England

Massimiliano Zampini
University of Oxford, Oxford, England
and University of Trento, Trento, Italy

and

Charles Spence
University of Oxford, Oxford, England

Participants made unspeeded temporal order judgments (TOJs) regarding which occurred first, an auditory or a visual target stimulus, when they were presented at a variety of different stimulus onset asynchronies. The target stimuli were presented either in isolation or positioned randomly among a stream of three synchronous audiovisual distractors. The largest just noticeable differences were reported when the targets were presented in the middle of the distractor stream. When the targets were presented at the beginning of the stream, performance was no worse than when the audiovisual targets were presented in isolation. Subsequent experiments revealed that performance improved somewhat when the position of the target was fixed or when the target was made physically distinctive from the distractors. These results show that audiovisual TOJs are impaired by the presence of audiovisual distractors and that this cost can be ameliorated by directing attention to the appropriate temporal position within the stimulus stream.

Our ability to extract meaningful information from the complex scenes of everyday life is enhanced by the integration of the sensory cues available to different sensory modalities (see, e.g., Stein & Meredith, 1993). The temporal synchrony and spatial coincidence of individual sensory stimuli are two of the key factors that modulate the perception of unified multisensory events (see, e.g., Calvert, Spence, & Stein, 2004; Driver & Spence, 2000; Slutsky & Recanzone, 2001). Previously, the multisensory perception of synchrony has normally been investigated using the temporal order judgment (TOJ) task (see Shore & Spence, 2005, and Spence, Shore, & Klein, 2001, for reviews). In a typical multisensory TOJ study, two stimuli from different sensory modalities are presented at various stimulus onset asynchronies (SOAs), and participants have to make unspeeded judgments regarding which sensory modality appeared to have been presented first.

Early multisensory TOJ studies using simple transitory stimuli (such as brief sound bursts and light flashes) suggested that auditory and visual stimuli needed to be separated by a minimum of 20 msec for people to be able to judge correctly which modality came first on 75% of the trials (the so-called just noticeable difference, or JND; see, e.g., Hirsh & Sherrick, 1961). However, it is important to note that in the majority of previous studies (including Hirsh & Sherrick's seminal study), the auditory and visual stimuli were presented from different spatial locations, and participants may therefore have used redundant spatial information to facilitate their TOJ responses (that is, participants may have judged which location came first rather than which modality came first; see Spence et al., 2001, and Zampini, Shore, & Spence, 2003). Subsequent studies in which such spatial confounds have been removed (by presenting the stimuli in different modalities from the same spatial location) have revealed that discrete pairs of auditory and visual stimuli need to be separated by approximately 60–70 msec in order for naive participants (i.e., those without extensive experience of psychophysical testing procedures) to judge accurately which modality was presented first (see, e.g., Zampini et al., 2003).

In order to investigate the temporal constraints on the multisensory perception of synchrony under more realistic conditions, one has to move away from the study of simple transitory stimuli of low informational content (like the stimuli typically used in the majority of previous TOJ studies; see, e.g., Hirsh & Sherrick, 1961; Zampini et al., 2003) to-

A. Vatakis, argiro.vatakis@psy.ox.ac.uk or argiro.vatakis@gmail.com

ward the use of more ecologically valid and complex stimuli such as, for example, speech, music, or object actions (see de Gelder & Bertelson, 2003; McGrath & Summerfield, 1985; Vatakis & Spence, 2006a, 2006b). However, the research published to date suggests that people's ability to detect the asynchrony of informationally rich audiovisual stimuli is very poor. For example, Dixon and Spitz (1980) reported that people noticed that a continuous stream of audiovisual speech was asynchronous only when the auditory stream led the visual stream by at least 131 msec, or when it lagged by 258 msec or more. More recently, Grant, van Wassenhove, and Poeppel (2004) reported that participants noticed the asynchrony in a continuous speech stream only when the speech sounds led the visual lip movements by at least 50 msec, or when they lagged by 220 msec or more.

The sensitivity to temporal asynchrony observed in the two studies cited above is much higher than that reported in studies that used simple auditory and visual stimuli (where JND values have typically ranged from 20–70 msec; see, e.g., Hirsh & Sherrick, 1961; Zampini et al., 2003). What accounts for such dramatic differences in temporal discrimination performance for simple versus complex audiovisual stimuli? In the present study, we address this question by investigating whether the presence of synchronous audiovisual distractors would affect TOJ performance for simple sound–light pairs. In our first experiment, asynchronous pairs of auditory and visual stimuli were presented either in isolation (distractor-absent blocks), as in most previous multisensory TOJ studies, or they were presented randomly among a stream of synchronous distractors (distractor-present blocks). If the presence of distractors adversely affects temporal discrimination performance, this might help to explain the marked differences in people's sensitivity to audiovisual synchrony revealed by previous researchers who used simple sound and light pairs (e.g., Hirsh & Sherrick, 1961; Zampini et al., 2003) versus those who used more complex speech stimuli (e.g., Dixon & Spitz, 1980; Grant et al., 2004; Vatakis & Spence, 2006a, 2006b; see also Fujisaki & Nishida, 2005).

EXPERIMENT 1

Method

Participants. Twenty-three participants (12 male and 11 female), between 20 and 32 years of age (mean age of 25) took part in the experiment. All of the participants reported having normal hearing and normal or corrected-to-normal vision. All were naive to the purpose of the experiment, and they varied in their previous experience of psychophysical testing procedures. The experiment took approximately 50 min to complete.

Apparatus and Stimuli. The experiment was conducted in a completely dark sound-attenuated booth. A loudspeaker cone (7 cm in diameter) was positioned centrally on a table 64 cm in front of the participant. A red light-emitting diode (LED), placed in the middle of the loudspeaker cone, was oriented so that when it was activated, it illuminated the surface of the loudspeaker cone. The auditory and visual stimuli were presented from exactly the same spatial location in order to avoid the spatial confound that can arise when stimuli from different sensory modalities are presented from different spatial locations (see Spence et al., 2001, and Zampini et al., 2003, on this point).

The auditory stimuli consisted of 8-msec bursts of white noise at 75 dB(A), as measured from the participant's ear position; the visual stimuli consisted of the illumination of the red LED for 8 msec. The target event always consisted of an asynchronous pairing of the audiovisual stimuli, whereas the distractors consisted of the simultaneous presentation of the auditory and visual stimuli. The target stimuli were either presented in isolation or positioned randomly among a stream of three synchronous audiovisual distractors. No specific attempt was made to match the intensities of the auditory and visual stimuli, which were both presented at a clearly suprathreshold level. White noise was presented continuously at 50 dB(A), as measured from the participants' ear position throughout the experiment from a loudspeaker positioned 14 cm directly above the target loudspeaker to mask any sounds made by participants.

Design. Two different block types were presented to participants in Experiment 1: In the distractor-absent blocks, the auditory and visual target stimuli were presented in isolation; in the distractor-present blocks, the asynchronous audiovisual target stimuli were presented together with three sequentially presented synchronous audiovisual distractors. The asynchronous target stimuli appeared randomly in Positions 1, 2, 3, or 4 of the stream (see Figure 1). The temporal interval between successive events varied randomly among four possible intervals (230, 255, 290, or 390 msec). The auditory and visual target stimuli were separated by one of eight possible SOAs (±546 msec, ±246 msec, ±141 msec, and ±66 msec; negative SOAs indicate that the auditory stimulus was presented first),[1] which varied from trial to trial according to the method of constant stimuli (Spence et al., 2001). Distractor-present and distractor-absent blocks were presented alternately with their order of presentation counterbalanced across participants. There were 32 trials in each of the four distractor-absent blocks and 128 trials in each of the four distractor-present blocks. Each of the four target positions was presented randomly four times at each of the eight SOAs in each block of trials. The participants completed three 15-trial practice blocks consisting of two distractor-absent blocks followed by one distractor-present block. The distractor-absent blocks were presented first to facilitate the acquisition of the task by participants.

Procedure. The participants received detailed verbal instructions prior to the start of the experiment, and they were allowed to ask for any clarification if they wished. The participants were informed that they would have to decide on each trial whether the auditory or the visual target stimulus had been presented first and respond by pressing one of two response keys placed on the table directly in front of them. The participants were instructed to press the sound key if they judged that the auditory target had appeared first and the light key if they judged that the visual target had appeared first. Participants had 6 sec after stimulus onset to respond (after which the trial was terminated); however, participants were instructed to respond only when confident of their decision. When participants responded "light first," feedback in the form of a 75-msec illumination of the red LED was provided; when they responded "sound first," a 75-msec burst of white noise was provided. The participants were informed that the feedback simply indicated which response they had made and did not indicate correctness. The participants were also informed that if they responded before stimulus presentation had been completed, or if they failed to make a response before the trial was terminated, error feedback, in the form of a 1,000-msec illumination of the red LED, would be presented. Such responses occurred on fewer than 3% of trials overall and were not analyzed. The first stimulus was presented 1,250 msec after the start of each trial. The participants were allowed to take a break between the blocks of experimental trials.

Results

The "vision first" responses (see Figure 2A) were converted to their equivalent *z*-scores assuming a cumulative normal distribution (cf. Finney, 1964). Best-fitting straight lines for each condition were calculated for each participant. Slope and intercept values were then derived from these straight lines. These two values were used to calculate the JND (JND = 0.675/slope, since +0.675 represents the 75% point and −0.675 represents the 25% point on the

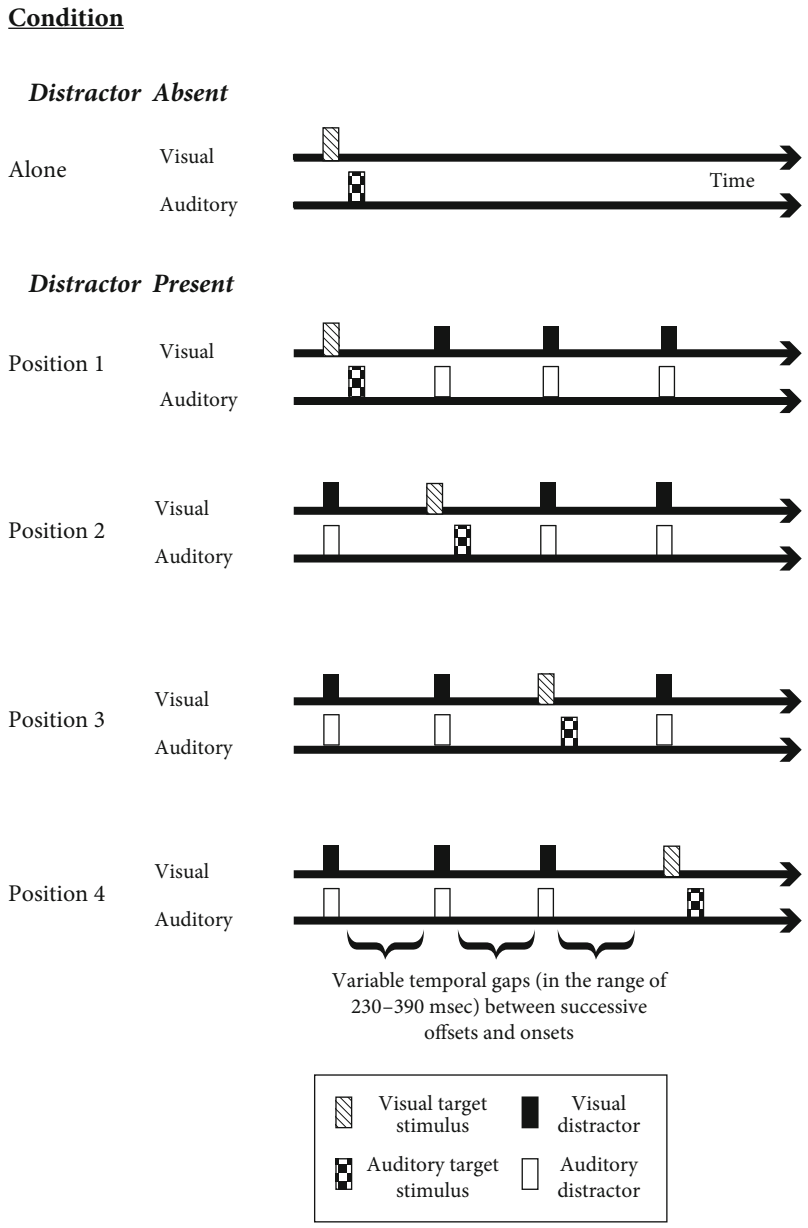

Figure 1. Schematic illustration of the five types of trials presented in Experiments 1–3. The asynchronous auditory and visual target stimuli could be presented either in isolation (distractor-absent condition) or randomly among three synchronous distractors (distractor-present condition, with the target being presented in positions 1, 2, 3, or 4, respectively). The four distractor-present conditions were presented randomly in Experiments 1 and 3 and in separate blocks of trials in Experiment 2. Visual target first (depicted) and auditory target first trials were presented equiprobably.

cumulative normal distribution) and the point of subjective simultaneity (PSS; PSS = −intercept/slope) values (see Coren, Ward, & Enns, 2004, for further details). The JND provides a standardized measure of the accuracy with which participants were able judge the temporal order of the auditory and visual target stimuli. The PSS indicates the amount of time by which one stimulus modality had to lead the other in order for synchrony to be perceived (i.e., for participants to make the "sound first" and "light first" responses equally often). For all of the analyses reported here, Bonferroni-corrected t tests (where $p < .05$ prior to correction) were used for all post hoc comparisons. The JND and PSS data were both analyzed using a one-way ANOVA with the factor of stimulus type (five levels: distractor absent, and distractor present in Positions 1, 2, 3, and 4). Data from 4 participants were removed from subsequent analysis because their PSS and/or JND values exceeded 600 msec (i.e., they fell outside the SOA range tested; cf. Spence et al., 2001, for similar exclusion criteria), indicating that these 4 participants were not able to perform the task.

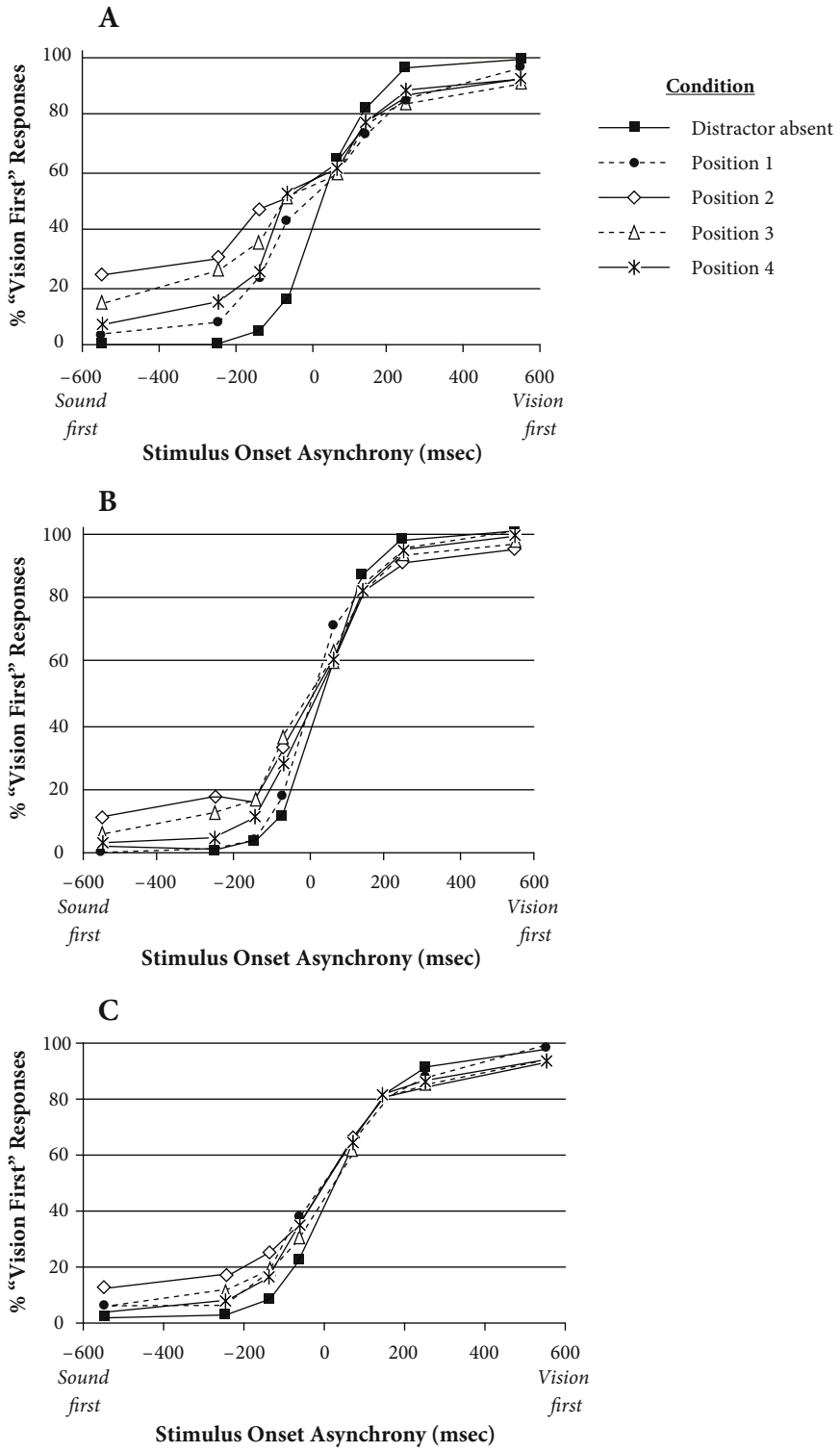

Figure 2. Mean percentage of "vision first" responses plotted as a function of stimulus onset asynchrony for each of the five conditions in Experiments 1 (A), 2 (B), and 3 (C).

Analysis of the JND data (see Figure 3A) revealed a significant main effect of stimulus type [$F(4,90) = 10.24$, $p < .01$]. Participants were significantly more sensitive to the temporal order of the auditory and visual target stimuli when the stimuli were presented in isolation (i.e., in the distractor-absent condition, mean JND = 108 msec) than when the target occurred in Position 2 ($M = 304$ msec) or 3 ($M = 248$ msec) of the distractor-present blocks

(both ps $< .01$, by t test pairwise comparisons). Participants were also significantly more sensitive to the temporal order of the asynchronous target stimuli when they were presented at the start of the distractor stream (i.e., in Position 1, $M = 160$ msec) than when they appeared in Position 2 or 3 (both ps $< .01$). Greater sensitivity in participants' responses was also observed when the asynchronous target was in the last position (i.e., Position 4; $M = 184$ msec) of the distractor stream than when the target was in Position 2 ($p < .01$). Although there were some numerical differences between the JND values reported in the distractor-absent blocks and the values reported when the audiovisual target stimuli were presented at either the start or the end of the distractor stream (i.e., in Position 1 or 4 of the distractor-present blocks, respectively; see Figure 3A), these differences failed to reach statistical significance.

A similar analysis of the PSS data (see Figure 3B) also revealed a significant main effect of stimulus type [$F(4,90) = 11.43$, $p < .01$]. When the audiovisual target stimuli were presented in isolation (i.e., the distractor-absent condition) or at the start of the stream in the distractor-present blocks (i.e., in Position 1), the onset of the visual target had to lead that of the auditory target by 38 msec and 6 msec, respectively, for the PSS to be achieved. This contrasts with the 165 msec and 64 msec by which the auditory stimulus had to lead when the target occurred in Positions 2 or 3, respectively, in the stream ($p < .01$ for both comparisons). Finally, the auditory stimulus had to lead the visual stimulus by a significantly greater interval for the PSS to be achieved when the target was presented in Position 2 ($M = 165$ msec) than when it was presented in Position 3 or 4 ($M = 25$ msec; $p < .05$; $p < .01$, respectively). With the exception of Positions 1 [$t(18) = 0.18$, $p = .86$] and 4 [$t(18) = -1.00$, $p = .33$], paired samples t tests revealed that all of the PSS values were significantly different from 0 msec (i.e., from veridical simultaneity).

Discussion

The results of Experiment 1 demonstrate that audiovisual TOJ performance was significantly impaired (i.e., JNDs were significantly higher) when the asynchronous audiovisual target stimuli were presented in the middle of a stream of synchronous audiovisual distractors (in Position 2 or 3 of the distractor-present blocks) than when they were presented at either the start or the end of the stream (in Position 1 or 4), or when they were presented in isolation (in the distractor-absent condition). Furthermore, the presence of the distractors also had a significant effect on the PSS. In particular, the visual stimulus had to be presented before the auditory stimulus in order for the PSS to be achieved when the target was presented in isolation or at the start of the stream (see Zampini et al., 2003, for similar results); for the middle positions (Positions 2 and 3), an auditory lead of 165 msec and 64 msec, respectively, was required.

The JND values obtained in Experiment 1 are noticeably higher than those typically observed in previous au-

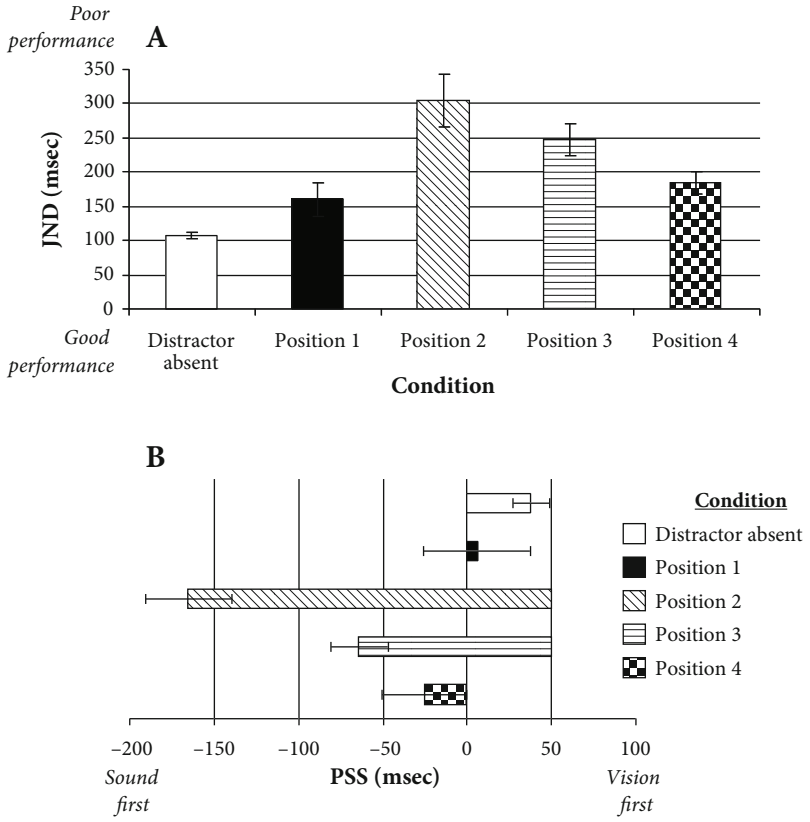

Figure 3. Mean JND (A) and PSS (B) values for the five conditions tested in Experiment 1. The error bars represent the standard errors of the means.

diovisual TOJ studies (see, e.g., Hirsh & Sherrick, 1961; Zampini et al., 2003). While this difference was expected in the distractor-present blocks (where the JNDs ranged from 160–304 msec), we were somewhat surprised to find JND values in excess of 100 msec in the distractor-absent blocks as well. The JND in the distractor-absent blocks (i.e., in the condition most similar to that used in previous audiovisual TOJ studies) is much larger (M = 106 msec) than that reported in many previous studies (e.g., the 20 msec reported in Hirsh & Sherrick's 1961 study, and the 60–70 msec reported in Zampini et al.'s 2003 study). One account for these differences might be the wider range of SOAs used in the present study (i.e., from −546 to +546 msec) as compared with the ±30-msec SOA range tested in Hirsh and Sherrick's study and the ±200-msec SOA range tested in Zampini et al.'s study. It seems plausible that the use of a wider range of SOAs in the present study might have resulted in participants' adopting a broader temporal focus for their attention than did participants in previous studies, resulting in poorer temporal discrimination performance (Navarra et al., 2005; Noesselt, Fendrich, Bonath, Tyll, & Heinze, 2005; but see Zampini, Guest, Shore, & Spence, 2005, on this point).[2]

One reason that the JNDs may have been so large in the distractor present blocks relates to the fact that the audiovisual target stimuli were presented randomly in each of the four positions in the stream on each trial. Consequently, the participants would have had no strategic reason to direct their attention to a particular point in the stimulus stream in advance. Previous research has shown that target discrimination performance can improve when targets appear at an expected point in time as compared with when they appear at an unexpected point in time (see, e.g., Correa, Sanabria, Spence, Tudela, & Lupiáñez, 2006; Lange & Röder, 2006; Nobre, 2001). Indeed, the fact that performance in Experiment 1 improved when the audiovisual target stimuli were presented at the end of the distractor stream, as compared with when they appeared at the two middle positions, may reflect the fact that participants, after having been presented with three synchronous distractors beforehand, were able to infer that the target would be in the last position and thus direct their attention temporally (Barnes & Jones, 2000; Riess-Jones, 2001).

We therefore thought it possible that performance in the distractor present blocks might improve if the position of the target in the stream were fixed and known to the participants in advance. We assessed this possibility in Experiment 2 by fixing the position of the asynchronous target stimuli within each block of trials: If participants were able to focus their attention on the appropriate temporal position within the stimulus stream, then one would expect to see an improvement in their temporal discrimination performance (and hence lower JNDs).

EXPERIMENT 2

Method

Participants. Thirteen new participants (5 male and 8 female) between 20 and 34 years of age (mean age of 24) took part in Experiment 2. One participant's data were removed from any subsequent analysis because of large PSS and/or JND values (cf. Spence et al., 2001).

Apparatus, Stimuli, Design, and Procedure. These were exactly the same as those used in Experiment 1 with the sole exception that the position of the asynchronous audiovisual target stimuli in the distractor-present blocks was fixed within each block of trials (rather than varying randomly from trial to trial, as was the case in Experiment 1). The participants were informed of the position of the asynchronous target stimuli within the distractor stream at the start of each distractor-present block. Two blocks of trials were presented for each position in the distractor-present stream. The order of presentation of each of the four distractor-present blocks was counterbalanced across participants with the sole restriction that the same target position was never presented in consecutive distractor present blocks.

Results and Discussion

Participants failed to respond before a trial was terminated on fewer than 1.5% of trials overall, and the data from these trials were not used in the following analyses. The proportion of "vision first" responses is highlighted in Figure 2B.

Analysis of the JND data (see Figure 4A) revealed a significant main effect of stimulus type [$F(4,55) = 5.93$, $p < .01$]. Participants were significantly more sensitive to the temporal order of the target stimuli in the distractor-absent blocks (M = 105 msec); they were also significantly more sensitive to the order of the target stimuli in the distractor-present blocks when the target appeared as the first item in the stream (M = 105 msec) than when it appeared in Position 2 (M = 205 msec; $p < .01$ for both comparisons), but not when it appeared in Position 3 (M = 161 msec) or 4 (M = 125 msec). The JND was significantly larger in blocks in which the target was the second or third item in the stream than in blocks in which the target occupied the last position in the stream (both ps = .02). None of the other comparisons reached significance. Analysis of the PSS data (see Figure 4B) revealed no main effect of stimulus type [$F(4,55) = 1.66$, $p = .17$].

Comparison of Performance in Experiments 1 and 2

We performed a combined analysis of the data from Experiments 1 and 2 to investigate whether advance knowledge of the position of the target in the distractor stream led to an improvement in accuracy of temporal discrimination responses. A between-experiments ANOVA on the JND data with the between-participants factor of Experiment (1 vs. 2) and the within-participants factor of stimulus type revealed a significant main effect of experiment [$F(1,29) = 8.31$, $p < .01$]. JNDs were significantly lower in Experiment 2 (M = 140 msec) than in Experiment 1 (M = 201 msec), showing that advance knowledge of the target position did indeed lead to a significant overall improvement in temporal discrimination performance. There was also a significant main effect of stimulus type [$F(4,116) = 16.82$, $p < .01$], with JNDs for discrimination of asynchronous targets presented in Positions 2 (M = 255 msec) and 3 (M = 204 msec) significantly higher than those for targets presented in Position 1 (M = 132 msec) or 4 (M = 155 msec), or when the distractors were absent (M = 106 msec). When the target was presented at the end of the stream, performance

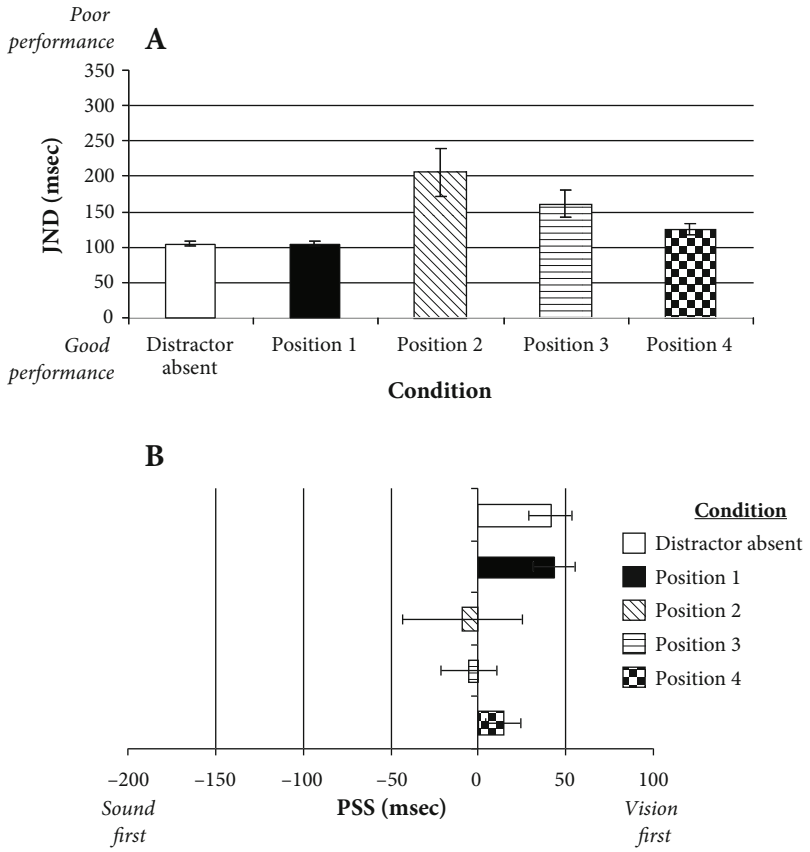

Figure 4. Mean JND (A) and PSS (B) values for the five conditions tested in Experiment 2. The error bars represent the standard errors of the means.

was also significantly worse than when the target was presented by itself. There was no interaction between experiment and stimulus type [$F(4,116) = 1.65, p = .17$].

A similar analysis of the PSS data revealed significant main effects of experiment [$F(1,29) = 8.00, p < .01$] and stimulus type [$F(4,116) = 12.50, p < .01$]. The main effect of experiment can be attributed to the fact that the auditory target had to lead the visual target by 42 msec, on average, for the PSS to be achieved in Experiment 1, whereas in Experiment 2, the visual stimulus had to lead by 17 msec, on average. Significant differences were also reported between the distractor-absent and Position 1 conditions (in which the visual stimulus had to lead by 40 msec and 25 msec, respectively, for the PSS to be achieved) and Positions 2 and 3 of the distractor-present conditions (in which auditory leads of 87 msec and 35 msec, respectively, were required). When the target was presented as the last item in the distractor stream, an auditory lead of 6 msec was required. There was also a significant interaction between experiment and stimulus type [$F(4,116) = 4.14, p < .01$], with PSS values in Experiment 2 differing significantly from those reported in Experiment 1 only for the middle two positions (i.e., 2 and 3) in the stream (both comparisons, $p < .01$).

Overall, the results of Experiment 2 highlight a significant improvement in TOJ performance relative to that seen in Experiment 1 when the position of the target was fixed within the stream of distractors. Presumably this improvement reflects the fact that participants were able to direct their attention temporally to the position of the target stimulus in the distractor stream in advance in Experiment 2 (see, e.g., Correa et al., 2006; Lange & Röder, 2006; Nobre, 2001), whereas they were unable to do this in Experiment 1 (except perhaps when the target appeared at the last position in the stream, when all uncertainty regarding the target position should have been removed). TOJ performance was better (i.e., JNDs were lower) when the target was presented at the start or end of the distractor stream than when it was presented in either of the middle two positions. Even though fixing the target position in Experiment 2 improved discrimination accuracy for audiovisual targets presented in the middle positions in the stream (i.e., Positions 2 and 3), performance in these conditions was still significantly worse than it was for the other two positions in the stream. The poor discrimination accuracy for target stimuli presented in the middle positions was observed despite the fact that participants knew in advance where in the distractor stream the target event would occur and presumably could have thus directed their temporal attention accordingly (Barnes & Jones, 2000; Riess-Jones, 2001).

Although the comparison of the results of Experiments 1 and 2 shows that fixing the position of a target in a distractor stream significantly improves temporal discrimination performance, one could argue that it does not unequivocally settle the question of whether the poor performance for

the middle target positions in the stream of distractors was due to poor temporal selection, task demands, or crowding (see, e.g., Chung, Levi, & Legge, 2001; Fujisaki & Nishida, 2005). In particular, we thought it possible that the similarity between the target and the distractors might have caused some confusion as to which was the target pair even when the position of the target pair was fixed within the distractor stream. Other possibilities are that participants may have inappropriately paired one of the target stimuli with one of the distractors or that the task might have been too demanding for the participants (given that they had not only to complete the TOJ task but also to determine the position of the target pair within the stream of distractors).

In order to discriminate between these various possibilities, we conducted a further experiment. In Experiment 3, we distinguished the target stimuli from the distractors by making the color of the visual target stimulus different from that of the distractor lights and by presenting the auditory target at a different frequency from the distractor sounds (cf. Fujisaki & Nishida, 2005). Thus, by reducing the likelihood that the target stimuli would be confused with the distractors, we were able to evaluate the confusion account of the results of Experiments 1 and 2. If performance for the middle target positions in the distractor-present stream in Experiment 3 was still significantly worse than it was for targets in the first and last positions of the stream, then this would suggest that crowding was responsible for the poor performance observed when the target was presented in the middle positions of the distractor stream.

Another possible account of the poor performance observed when the target pair occupied either of the middle positions in the distractor stream in Experiments 1 and 2 may be related to the phenomenon of intramodal perceptual grouping (see, e.g., Sanabria, Soto-Faraco, Chan, & Spence, 2004; Spence, Sanabria, & Soto-Faraco, in press; Wertheimer, 1938). That is, unimodal perceptual grouping may have occurred due to the physical similarity between the target and distractor stimuli (in terms of the duration of the stimuli and the matching of the frequency and color of the target with the distractor stimuli). Grouping may also have occurred because equal numbers of auditory and visual stimuli were presented, and the auditory and visual stimulus pairs were presented close together in both space and time. However, any such unimodal perceptual grouping of the target with the distractors (in either audition or vision) should be interrupted by presenting target and distractor in different colors and frequencies (see Spence et al., in press; Vroomen & de Gelder, 2000). Therefore, if temporal discrimination performance is still impaired for the middle target positions in Experiment 3, then the grouping account cannot successfully explain the pattern of results in the two experiments reported so far.

EXPERIMENT 3

Method

Participants. Twenty-four new participants (9 male and 15 female) between 20 and 32 years of age (mean age of 25) took part in Experiment 3.

Apparatus, Stimuli, Design, and Procedure. These were exactly the same as those used in Experiment 1 with the sole exception that the auditory target was presented at a different frequency (1500 Hz) from the synchronous distractor tones (500 Hz), and the visual target consisted of the brief illumination of a red LED, whereas the synchronous visual distractors were green (the result of a bicolor LED). The participants were instructed to report whether the higher frequency target sound was presented before or after the red target light while ignoring the lower frequency tones presented in synchrony with the green lights.[3]

Results and Discussion

The proportion of "vision first" responses is highlighted in Figure 2C. A one-way ANOVA performed on the JND data (see Figure 5A) derived from these psychometric functions revealed a significant main effect of stimulus type [$F(4,119) = 3.50$, $p = .01$]. Participants were significantly more sensitive to the temporal order of the target stimuli in the distractor-absent blocks ($M = 116$ msec) and when the target appeared as the first item in the stream ($M = 127$ msec; $p = .01$) than in the distractor-present blocks when the target appeared in Position 2 ($M = 225$ msec; $p = .03$), but not when it appeared in Position 3 ($M = 163$ msec) or 4 ($M = 145$ msec). None of the other comparisons reached statistical significance. A similar analysis of the PSS data (see Figure 5B) revealed no main effect of stimulus type [$F(4,119) = 1.37$, $p = .25$].

Comparison of Performance in Experiments 1 and 3 and Experiments 2 and 3

We performed a combined analysis of the data from Experiments 1 and 3 in order to investigate whether the physical distinctiveness of the target pair in the distractor stream introduced in Experiment 3 led to an improvement in the accuracy of temporal discrimination performance relative to that seen when target position was uncertain (Experiment 1). An ANOVA on the JND data with the between-participants factor of experiment (1 vs. 3) and the within-participants factor of stimulus type revealed a significant main effect of experiment [$F(1,41) = 3.61$, $p \leq .05$]. This result demonstrates that the distinctiveness of the target pair in Experiment 3 led to a significant improvement in temporal discrimination performance (JND = 155 msec) over performance when the target pair was not physically distinctive (Experiment 1; JND = 201 msec). This result adds further support to the view that the poor temporal discrimination performance reported for targets presented in the middle of the distractor stream in the distractor-present blocks of Experiment 1 may be attributed to uncertainty on the part of the participants about the temporal position of the target stimuli in the distractor stream (Barnes & Jones, 2000; Riess-Jones, 2001). There was also a significant main effect of stimulus type [$F(4,164) = 20.64$, $p < .01$], with significantly larger JNDs for asynchronous targets presented in Positions 2 ($M = 265$ msec) and 3 ($M = 205$ msec) than for targets presented in Position 1 ($M = 144$ msec) or 4 ($M = 165$ msec) or for targets presented when the distractors were absent ($M = 112$ msec). When the target pair was presented at the end of the distractor stream (i.e., in Position 4), performance was also significantly worse than

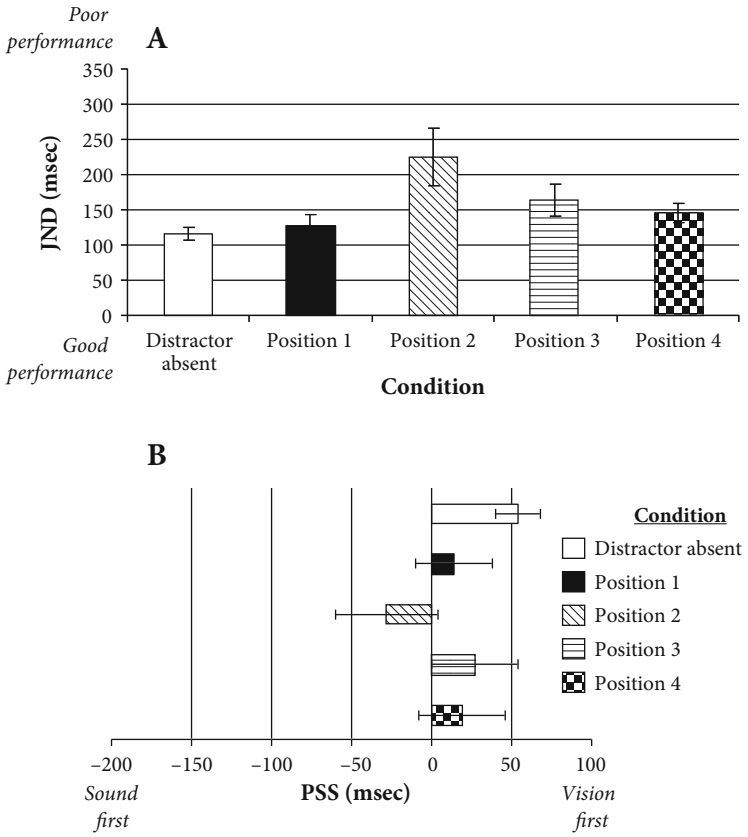

Figure 5. Mean JND (A) and PSS (B) values for the five conditions tested in Experiment 3. The error bars represent the standard errors of the means.

when the target pair was presented by itself. There was no interaction between experiment and stimulus type [$F(4,164) = 2.13$, $p = .1$].

A similar analysis of the PSS data revealed no main effect of experiment [$F(1,41) = 5.01$, $p < .05$]. However, a significant main effect of stimulus type [$F(4,164) = 16.53$, $p < .01$], and a significant experiment × stimulus type interaction were obtained [$F(4,164) = 4.36$, $p < .01$]. Significant differences were reported between the distractor-absent and distractor-present conditions in Positions 2, 3, and 4 (in which the auditory stimulus had to lead by 97 msec, 19 msec, and 3 msec, respectively, for the PSS to be achieved). In addition, significant differences were reported between Position 2 and Positions 1 (in which visual leads of 10 msec were required), 3, and 4 of the distractor-present conditions. There was also a significant interaction between experiment and stimulus type, with PSS values in Experiment 3 differing significantly from those reported in Experiment 1 only for the middle two positions in the stream (both comparisons, $p < .01$).

Given these results, it should come as no surprise that a similar between-experiments analysis of the data from Experiments 2 and 3 revealed no main effect of experiment [$F(1,34) = 0.34$, $p = .56$] or any interaction between experiment and stimulus type [$F(4,136) = 0.11$, $p = .98$]. These latter results demonstrate that the physical distinctiveness of the target pair introduced in Experiment 3 did not lead to a significant improvement in temporal discrimination performance over and above that attributable to fixing the position of the target in advance. Both experimental manipulations seem to have provided the participants with sufficient information with which to identify the target successfully. This comparison also shows that the poor performance in the distractor-present blocks cannot be attributed solely to the distractors' facilitating unimodal auditory and/or visual perceptual grouping, since changing the color and frequency of the target pair should have eliminated any tendency to group the target stimuli with the distractors (see, e.g., Spence et al., in press; Vroomen & de Gelder, 2000). A similar between-experiments analysis of the PSS data also revealed no main effect of experiment [$F(1,34) = 0.50$, $p = .99$], or any interaction between experiment and stimulus type [$F(4,136) = 0.79$, $p = .54$].

Overall, the comparisons of Experiments 1 and 3 and Experiments 2 and 3 revealed that there was significant improvement in performance when the target pair was made distinctive (either by changing its physical properties or by directing the participant to the position of the target pair) from the distractor pairs in the stream. Making the target physically distinctive from the distractors was no more effective in improving temporal discrimination performance than making the target position predictable (Fujisaki & Nishida, 2005). Such an outcome verifies the

fact that crowding persists even if the observer knows which stimulus constitutes the target (Montaser-Kouhsari & Rajimehr, 2005).

GENERAL DISCUSSION

Taken together, the results of the three experiments reported here show that people find it particularly difficult to discriminate the temporal order of asynchronous audiovisual target events when they are presented in the middle of a stream of synchronous audiovisual distractors. JNDs were significantly lower when the target stimuli were presented in isolation or at the start or end of a stream of synchronous audiovisual distractors than when they were presented in the middle of the stream (in Positions 2 and 3). Performance for targets presented in the middle of the stream improved when the position of the target within the distractor stream was fixed (Experiment 2) rather than when it varied randomly from trial to trial (Experiment 1). Similarly, temporal discrimination accuracy for the targets in the middle of the stream improved significantly when the target and distractor stimuli were made physically different from the distractors (Experiment 3), but accuracy was still poor compared with accuracy for the other positions in the stream (i.e., compared with Positions 1 and 4). Finally, even though the JNDs were numerically somewhat smaller in the distractor-absent condition than when the asynchronous targets were presented at the start of the audiovisual stream (in Position 1) in the distractor-present condition, this difference failed to reach statistical significance in any of the experiments (or in any of the between-experiments comparisons).

Can the detrimental effects of the presence of a stream of synchronous audiovisual distractors on TOJ performance be explained in terms of a crowding effect? Crowding (or lateral masking) typically refers to the decreased discrimination, detection, or recognition performance observed when a visual target is presented among spatially adjacent but nonoverlapping distractor stimuli (see, e.g., Chung et al., 2001). Similarly, auditory research has shown that target discrimination performance sometimes deteriorates when a target sound is embedded within a temporal sequence (or spatial array) of masking auditory stimuli (see, e.g., Chan, Merrifield, & Spence, 2005; Kidd, Mason, Rohtla, & Deliwala, 1998; Leek & Watson, 1984). Especially relevant to the present research is a review of studies by Yost and Watson (1987) in which they found better performance (in terms of accuracy) for auditory targets presented at the start or end of an auditory sequence than for targets presented in the middle positions of the sequence. The performance decrement reported in previous unimodal crowding studies has typically been attributed to the fact that participants had to attend to the stimulus stream and extract salient information regarding the target stimulus while at the same time trying to ignore (or suppress) the processing of the distractor stimuli. If the distractors are presented too close to the target in either space or time, then the distractors' crowding the target may impair selection and thus result in a performance decrement.

In the present study, we thought it possible that the distractors occurring before and after the target when the target was presented in the middle (i.e., Positions 2 or 3) of the stream of identical synchronous audiovisual distractors may have impaired temporal discrimination performance because of crowding. In our study, the temporal gap between the offset of the preceding synchronous distractor and the onset of the first element of the asynchronous target pair was in the range of 222–382 msec. A similar temporal gap also separated the offset of the second target element and the onset of the next distractor in the stream. The crowding account would certainly predict poorer performance (i.e., higher JNDs) for targets presented in the middle of a stream of distractors than for those presented at the beginning or end of the stream, just as we observed. In addition, the crowding account may also explain the poorer performance observed when the target was presented at the end of the stream (i.e., in Position 4) compared with when it was in Position 1, since recent studies have not limited the effects of crowding to only the middle positions of a sequence of stimuli (see Pelli, Palomares, & Majaj, 2004). We therefore believe that our results may provide the first evidence of crowding in a multisensory setting (as compared with previous studies of unimodal crowding; e.g., those by Chung et al., 2001; Kidd et al., 1998; Leek & Watson, 1984); our results also provide the first evidence of the effects of crowding specifically on temporal discrimination performance. Even though the mechanisms underlying crowding are still not fully understood (see, e.g., Cavanagh, 2001), our results nevertheless demonstrate that the effect of crowding must be different from that of grouping (since no support for a grouping account was found in the comparison between the results of Experiments 1 and 2). Finally, our results also show that whatever the origin of this crowding mechanism may be, its effects can nevertheless be reduced significantly (although not completely eliminated) by both top-down (Experiment 2) and bottom-up (Experiment 3) segregation.

Previous studies have shown that people can selectively orient their attention to a particular point in time (i.e., within a trial) and that this can improve performance (although note that none of the studies reported to date have looked at streams of stimuli, but have typically looked at abrupt, discrete stimulus presentation instead; see, e.g., Correa et al., 2006; Lange & Röder, 2006; Nobre, 2001). Our results therefore both confirm and extend these previous findings by showing that temporal attention may play an important role in the perception of multisensory temporal events. In future research, it would be of interest to vary the temporal separation between successive items in the stream in order to determine just how large an interval is needed between them before observers are able to ignore the distractors successfully when making audiovisual TOJs.

Recent research by Fujisaki and Nishida (2005) has shown that the temporal perception of multisensory events is also affected by changes in the temporal frequency at which stimuli are presented. Specifically, Fujisaki and Nishida reported a series of experiments showing that

participants' discrimination performance in a synchrony judgment task deteriorated as the temporal modulation frequency increased above 2–4 Hz using periodic pulse trains of stimuli. The authors suggested that this decrement in performance was due to the difficulty participants had in separating the salient temporal features from the other rapid repetitive signals; such separation would have assisted them in detecting the audiovisual synchrony of a specific target event that they were interested in. Fujisaki and Nishida's findings therefore support the hypothesis that the temporal processing of audiovisual signals is governed by the slow and attention-demanding processing of the various signal attributes and the consequent computation of the temporal relationships of the different signals as a function of the saliency of the signal features.

In the present study, the visual target had to be presented before the auditory target in order for synchrony to be perceived (i.e., for the PSS to be achieved) for asynchronous audiovisual pairs presented in the first position (see Zampini et al., 2003, for similar results) and last position of the stream, whereas auditory leads were required for the two middle positions (i.e., in the conditions that participants found most difficult). Large individual differences in the PSS were also observed between participants in all three of the experiments reported here (see the error bars in Figures 3B, 4B, and 5B), a trend that has also been observed in previous TOJ studies (see, e.g., Stone et al., 2001; see also Mollon & Perkins, 1996). The greatest interparticipant variability was noted when the auditory target was presented before the visual target (e.g., see the left side of the graphs shown in Figure 2), which may be due to the fact that we are naturally biased toward events that are first presented visually (see, e.g., Fujisaki, Shimojo, Kashino, & Nishida, 2004; Grant, Greenberg, Poeppel, & van Wassenhove, 2004; see also Spence & Squire, 2003). Such differences were not unexpected given previous reports regarding the variability of the PSS, both between individuals (see, e.g., Arnold, Johnston, & Nishida, 2005; Mollon & Perkins, 1996; Stone et al., 2001) and between different studies (see Spence et al., 2001, for a review). Understanding the factors that contribute to determining the PSS, both between individuals and between studies, remains an important issue for future research (cf. Spence & Squire, 2003).

In conclusion, the novel paradigm used in the present study allowed us to examine audiovisual temporal perception for audiovisual targets presented in a continuous stream of distractors. Audiovisual temporal discrimination performance was shown to be significantly worse for asynchronous target stimuli presented in the middle of a stream of distractors than for targets occurring either in isolation or at the start or end of the stream. Performance improved when the asynchronous target was made physically distinctive (in color and frequency) from the distractors or when its position was fixed within the stream. Our paradigm might be extended further in future studies to investigate temporal perception under more ecologically valid and informationally rich stimulus settings. It would be interesting, for example, to investigate how sensitive we are to asynchrony for audiovisual target words or notes when they are presented in isolation versus as part of a continuous audiovisual stream of speech or music, respectively (cf. Stolz, 1999; Vatakis & Spence, 2006a, 2006b). Such investigations will be necessary if one wishes to move away from simple stimuli (such as light flashes and/or sound bursts) to stimuli that apply more directly to our everyday experience (e.g., International Telecommunication Union, 1998; Reeves & Voelker, 1993). It is only by studying the temporal aspects of perception for more complex, informationally rich stimuli that we will be better able to understand the perception of multisensory synchrony and the factors that modulate our everyday perception of unified multisensory events.

AUTHOR NOTE

A.V. was supported by a Newton Abraham Studentship from the Medical Sciences Division, University of Oxford. Correspondence regarding this article should be addressed to A. Vatakis, Department of Experimental Psychology, University of Oxford, South Parks Road, Oxford OX1 3UD, England (e-mail: argiro.vatakis@psy.ox.ac.uk or argiro.vatakis@gmail.com).

REFERENCES

ARNOLD, D. H., JOHNSTON, A., & NISHIDA, S. (2005). Timing sight and sound. *Vision Research*, **45**, 1275-1284.

BARNES, R., & JONES, M. R. (2000). Expectancy, attention, and time. *Cognitive Psychology*, **41**, 254-311.

CALVERT, G. A., SPENCE, C., & STEIN, B. E. (EDS.). (2004). *The handbook of multisensory processes*. Cambridge, MA: MIT Press.

CAVANAGH, P. (2001). Seeing the forest but not the trees. *Nature Neuroscience*, **4**, 673-674.

CHAN, J. S., MERRIFIELD, K., & SPENCE, C. (2005). Auditory spatial attention assessed in a flanker interference task. *Acta Acustica*, **91**, 554-563.

CHUNG, S. T. L., LEVI, D. M., & LEGGE, G. E. (2001). Spatial-frequency and contrast properties of crowding. *Vision Research*, **41**, 1833-1850.

COREN, S., WARD, L. M., & ENNS, J. T. (2004). *Sensation and perception* (6th ed.). Hoboken, NJ: Wiley.

CORREA, A., SANABRIA, D., SPENCE, C., TUDELA, P., & LUPIÁÑEZ, J. (2006). Selective temporal attention enhances the temporal resolution of visual perception: Evidence from a temporal order judgment task. *Brain Research*, **1070**, 202-205.

DE GELDER, B., & BERTELSON, P. (2003). Multisensory integration, perception and ecological validity. *Trends in Cognitive Sciences*, **7**, 460-467.

DIXON, N. F., & SPITZ, L. (1980). The detection of auditory visual desynchrony. *Perception*, **9**, 719-721.

DRIVER, J., & SPENCE, C. (2000). Multisensory perception: Beyond modularity and convergence. *Current Biology*, **10**, R731-R735.

FINNEY, D. J. (1964). *Probit analysis: A statistical treatment of the sigmoid response curve* (2nd ed.). London: Cambridge University Press.

FUJISAKI, W., & NISHIDA, S. (2005). Temporal frequency characteristics of synchrony–asynchrony discrimination of audio-visual signals. *Experimental Brain Research*, **166**, 455-464.

FUJISAKI, W., SHIMOJO, S., KASHINO, M., & NISHIDA, S. (2004). Recalibration of audiovisual simultaneity. *Nature Neuroscience*, **7**, 773-778.

GRANT, K. W., GREENBERG, S., POEPPEL, D., & VAN WASSENHOVE, V. (2004). Effects of spectro-temporal asynchrony in auditory and auditory-visual speech processing. *Seminars in Hearing*, **25**, 241-255.

GRANT, K. W., VAN WASSENHOVE, V., & POEPPEL, D. (2004). Detection of auditory (cross-spectral) and auditory-visual (cross-modal) synchrony. *Speech Communication*, **44**, 43-53.

HIRSH, I. J., & SHERRICK, C. E., JR. (1961). Perceived order in different sense modalities. *Journal of Experimental Psychology*, **62**, 423-432.

INTERNATIONAL TELECOMMUNICATIONS UNION (1998). *Relative timing*

of sound and vision for broadcasting (ITU-R BT.1359-1). Geneva: Author.

KIDD, G., JR., MASON, C. R., ROHTLA, T. L., & DELIWALA, P. S. (1998). Release from masking due to spatial separation of sources in the identification of nonspeech auditory patterns. *Journal of the Acoustical Society of America*, **104**, 422-431.

LANGE, K., & RÖDER, B. (2006). Orienting attention to points in time improves stimulus processing both within and across modalities. *Journal of Cognitive Neuroscience*, **18**, 715-729.

LEEK, M. R., & WATSON, C. S. (1984). Learning to detect auditory pattern components. *Journal of the Acoustical Society of America*, **76**, 1037-1044.

MCGRATH, M., & SUMMERFIELD, Q. (1985). Intermodal timing relations and audio-visual speech recognition by normal-hearing adults. *Journal of the Acoustical Society of America*, **77**, 678-685.

MOLLON, J. D., & PERKINS, A. J. (1996). Errors of judgement at Greenwich in 1796. *Nature*, **380**, 101-102.

MONTASER-KOUHSARI, L., & RAJIMEHR, R. (2005). Subliminal attentional modulation in crowding condition. *Vision Research*, **45**, 839-844.

NAVARRA, J., VATAKIS, A., ZAMPINI, M., SOTO-FARACO, S., HUMPHREYS, W., & SPENCE, C. (2005). Exposure to asynchronous audiovisual speech extends the temporal window for audiovisual integration. *Cognitive Brain Research*, **25**, 499-507.

NOBRE, A. C. (2001). Orienting attention to instants in time. *Neuropsychologia*, **39**, 1317-1328.

NOESSELT, T., FENDRICH, R., BONATH, B., TYLL, S., & HEINZE, H. J. (2005). Closer in time when farther in space—Spatial factors in audiovisual temporal integration. *Cognitive Brain Research*, **25**, 443-458.

PELLI, D. G., PALOMARES, M., & MAJAJ, N. J. (2004). Crowding is unlike ordinary masking: Distinguishing feature integration from detection. *Journal of Vision*, **4**, 1136-1169.

REEVES, B., & VOELKER, D. (1993). *Effects of audio–video asynchrony on viewer's memory, evaluation of content and detection ability*. Los Gatos, CA: Pixel Instruments.

RIESS-JONES, M. (2001). Temporal expectancies, capture, and timing in auditory sequences. In C. Folk & B. Gibson (Eds.), *Attraction, distraction, and action: Multiple perspectives on attentional capture* (pp. 191-229). New York: Elsevier Science.

SANABRIA, D., SOTO-FARACO, S., CHAN, J. S., & SPENCE, C. (2004). When does visual perceptual grouping affect multisensory integration? *Cognitive, Affective, & Behavioral Neuroscience*, **4**, 218-229.

SHORE, D. I., & SPENCE, C. (2005). Prior entry. In L. Itti, G. Rees, & J. K. Tsotsos (Eds.), *Neurobiology of attention* (pp. 89-95). San Diego: Elsevier, Academic Press.

SLUTSKY, D. A., & RECANZONE, G. H. (2001). Temporal and spatial dependency of the ventriloquism effect. *NeuroReport*, **12**, 7-10.

SPENCE, C., SANABRIA, D., & SOTO-FARACO, S. (in press). Intersensory Gestalten and crossmodal scene perception. In K. Noguchi (Ed.), *The psychology of beauty and Kansei: New horizons of Gestalt perception.*

SPENCE, C., SHORE, D. I., & KLEIN, R. M. (2001). Multisensory prior entry. *Journal of Experimental Psychology: General*, **130**, 799-832.

SPENCE, C., & SQUIRE, S. (2003). Multisensory integration: Maintaining the perception of synchrony. *Current Biology*, **13**, R519-R521.

STEIN, B. E., & MEREDITH, M. A. (1993). *The merging of the senses.* Cambridge, MA: MIT Press.

STOLZ, J. A. (1999). Word recognition and temporal order judgments: Semantics turns back the clock. *Canadian Journal of Experimental Psychology*, **53**, 316-322.

STONE, J. V., HUNKIN, N. M., PORRILL, J., WOOD, R., KEELER, V., BEANLAND, M., ET AL. (2001). When is now? Perception of simultaneity. *Proceedings of the Royal Society of London: Series B*, **268**, 31-38.

VATAKIS, A., & SPENCE, C. (2006a). Audiovisual synchrony perception for music, speech, and object actions. *Brain Research*, **1111**, 134-142.

VATAKIS, A., & SPENCE, C. (2006b). Audiovisual synchrony perception for speech and music assessed using a temporal order judgment task. *Neuroscience Letters*, **393**, 40-44.

VROOMEN, J., & DE GELDER, B. (2000). Sound enhances visual perception: Cross-modal effects of auditory organization on vision. *Journal of Experimental Psychology: Human Perception & Performance*, **26**, 1583-1590.

WERTHEIMER, M. (1938). Laws of organization in perceptual forms. In W. D. Ellis (Ed.), *A source book of Gestalt psychology* (pp. 71-88). London: Kegan Paul, Trench, Trubner.

YOST, W. A., & WATSON, C. S. (EDS.) (1987). *Auditory processing of complex sounds*. Hillsdale, NJ: Erlbaum.

ZAMPINI, M., GUEST, S., SHORE, D. I., & SPENCE, C. (2005). Audio–visual simultaneity judgments. *Perception & Psychophysics*, **67**, 531-544.

ZAMPINI, M., SHORE, D. I., & SPENCE, C. (2003). Audiovisual temporal order judgments. *Experimental Brain Research*, **152**, 198-210.

NOTES

1. In a pilot study, we used SOAs of ±190, ±90, ±55, and ±30 msec. Participants' performance suggested that the task was too difficult at these SOA levels, so higher SOA values were used in the experiments reported here.

2. In order to explore whether the larger JNDs obtained in the distractor-absent blocks of Experiment 1 (as compared to those observed in previous studies; i.e., Hirsh & Sherrick, 1961, and Zampini et al., 2003) were attributable to the larger intervals used in Experiment 1 or to the alternation between distractor-absent and distractor-present blocks, we conducted a follow-up study. Five new participants were presented with four blocks of distractor-absent trials in which only a single pair of auditory and visual target stimuli appeared, as in the majority of previous audiovisual TOJ studies (see, e.g., Hirsh & Sherrick, 1961, and Zampini et al., 2003); the participants had to make TOJs regarding which modality, audition or vision, appeared to have been presented first. The results (mean JND = 120 msec) were very similar to those obtained in the distractor-absent blocks of Experiment 1 (mean JND = 107 msec), showing that the larger JND values in the present study presumably reflect our use of larger SOAs rather than the alternation between distractor-present and distractor-absent blocks.

3. In order to ensure that the participants were indeed able to correctly identify the position of the target stimuli in the distractor sequence, we conducted a follow-up study in which 5 participants were presented with two distractor-present blocks and had to identify the position of the target by pressing buttons 1, 2, 3, or 4 on a standard computer keyboard. The results showed that participants were able to identify the position of the target correctly regardless of whether it appeared in Positions 1, 2, 3, or 4 in the stimulus stream, with participants responding correctly on more than 97% of the trials in each of the 4 positions.

(Manuscript received April 21, 2005;
revision accepted for publication April 19, 2006.)