RUNNING HEAD: AUDIOVISUAL SPEECH INTEGRATION AND SPATIAL ATTENTION

Searching for audiovisual correspondence in multiple speaker scenarios

Agnès Alsius[1] and Salvador Soto-Faraco[2,3]

1. Departament de Psicologia Bàsica, Universitat de Barcelona, Barcelona, Spain

2. Departament de Tecnologies de la Informació i les Comunicacions, Universitat Pompeu Fabra, Barcelona, Spain

3. Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain

Character count abstract, main text and references (max 35000, inc. spaces): 35,236

Number of text pages: 31

Number of figures: 4

Contact address: Agnès Alsius Rancé
Departament de Psicologia Bàsica
Universitat de Barcelona
Pg Vall d'Hebrón, 171
08035 Barcelona, SPAIN
Contact e-mail: aalsius@gmail.com
Telephone: +34 93. 312.51.44
Fax: +34 93. 402.13.63

KEYWORDS: MULTISENSORY INTEGRATION; AUDIOVISUAL SPEECH PERCEPTION; SPATIAL ATTENTION; VISUAL SEARCH; AUDITORY SEARCH

Abstract

A critical question in multisensory processing is how the constant information flow that arrives to our different senses is organized in coherent representations. Some authors claim that pre-attentive detection of inter-sensory correlations supports crossmodal binding, whereas other findings indicate that attention plays a crucial role. We used visual and auditory search tasks for speaking faces to address the role of selective spatial attention in audiovisual binding. Search efficiency amongst faces for the match with a voice declined with the number of faces being monitored concurrently, consistent with an attentive search mechanism. In contrast, search amongst auditory speech streams for the match with a face was independent of the number of streams being monitored concurrently, as long as localization was not required. We suggest that the fundamental differences in the way in which auditory and visual information is encoded play a limiting role in crossmodal binding. Based on these unisensory limitations, we provide a unified explanation for several previous apparently contradictory findings.

Introduction

In order to orchestrate the vast amount of information arriving from the sensory afferents, our perceptual system must detect valid co-occurrences across senses (i.e., information originating from the same event), and discard spurious coincidences (i.e., inputs from causally independent origins) in what has been called the crossmodal pairing problem (e.g., Bertelson and de Gelder 2004; Welch 1999; Welch and Warren 1980). However, how this inter-sensory pairing process is carried out is not altogether clear. An unresolved question is whether inter-sensory correlations are detected and bound pre-attentively or else, if they require selective spatial attention.

Fujisaki et al. (2006; see also Van de Par and Kohlrausch 2004) examined whether the detection of audiovisual temporal synchrony amongst multiple objects followed a parallel or a serial search pattern[1]. They found that the efficiency in detecting a visual target that flickered in synchrony with a beeping sound was inversely proportional to the number of visual (unsynchronized) distractors. This suggests that detection of audio-visual synchrony requires attention, in contrast with previous claims that audiovisual binding proceeds automatically (Bertelson et al. 2000a, b; Colin et al. 2002; Driver 1996; Spence and Santangelo 2009; Vroomen et al. 2001), which would predict that the unique visual element in the display that matched the sound should have popped out from the distractors.

[1] Parallel search is inferred when the reaction times for detecting a target amongst distractors does not increase with display size (i.e., the number of distractors), a result which leads to the interpretation that the target can be identified in the absence of selective attention. On the other hand, attentive serial search is typically characterized by a linear increase in reaction times to the target as a function of the number of elements in the search array, revealing a search mechanism that presumably involves checking each display element individually until the target is found. Some studies have questioned a strict dichotomy between parallel and serial search, favouring, instead, a continuum of attentional requirements (Duncan and Humphreys 1989; Joseph et al. 1997; McElree and Carrasco 1999). Furthermore, it has been shown that both serial and parallel mechanisms can work together to identify objects in a scene (Bichot et al. 2005). For this reason, we will not use these theoretically loaded terms (i.e., serial, parallel) in the present manuscript, and will merely describe the results in terms of search efficiency (Wolfe 2003).

Previous studies have shown that a spatially non-informative auditory event can indeed guide attention towards the location of a semantically (Iordanescu et al. 2008) or temporally matching (Van der Burg et al. 2008 a, b; 2010) visual object in complex scenes. In Van der Burg et al. (2008 a) study, participants were asked to find a horizontal or vertical bar amongst a varying number of oblique bars, a paradigm that usually leads to steep search slopes indicating a serial process (RTs ranged from 3 to 7 sec, for displays containing 24 to 48 elements). Display items (both target and distractors) changed color (green to/from red) at unpredictable times. Critically, when a spatially uninformative beep was paired with the colour change of the target, despite the overall long search times (>2sec) search slopes became nearly flat (although note that they never reached the values commonly attributed to a strict parallel search). This suggests that the sound boosted the saliency of the concurrent visual event and helped overcome limitations in visual spatial attention. The diverging outcomes between Fujisaki et al. and Van der Burg et al. studies may root in the varying perceptual load associated to the accessory stimulus (Talsma et al. 2010; Van der Burg et al. 2010; see also Lavie 2005). In studies where sounds failed to aid visual search, both the visual (search modality) and the auditory (accessory modality) events unfolded at much higher rates (i.e., up to 40 Hz; Fujisaki et al. 2006; Van de Par and Kohlrausch 2004).

Here, we addressed the role of attention in audiovisual congruence detection for speech and expose potential limitations imposed by unimodal attention. Audiovisual speech is rich in temporal and spectral information over time and yet, unlike stimuli (i.e., beeps and flashes) used in past studies, is a biologically relevant over learned stimulus for humans. Observers may be especially sensitive to the inter-sensory correlations and predictability between the speakers' face and speech sounds (e.g., Kamachi et al. 2006; Pickering and Garrod 2006; Van Wassenhove et al. 2005).

Interestingly, although audiovisual speech integration is one of the most often cited illustrations of the automatic nature of multisensory binding (Colin et al. 2002; McGurk and MacDonald 1976) the few studies addressing whether it requires selective attention have yielded mixed results (Alsius et al. 2005 2007; Andersen et al. 2008; Driver 1996; Fairhall and Macaluso 2009; Soto-Faraco et al. 2004; see Navarra et al. 2009 for a review). Fairhall and Macaluso (2009) found that directing covert attention to a talking face matching a concurrent speech stream, as opposed to attending a non-matching face within the same dislpay, was critical for the activation of brain regions reflecting audiovisual integration. In contrast, Driver (1996) showed that selective listening of one out of two spatially separated speech streams worsened when the lip movements corresponding to the target speech stream were displayed near the distractor sounds. The interpretation was that the matching (target) auditory stream was ventriloquized toward the location of the seen talker face, thus making the spatial segregation of targets from distractors more difficult. This finding suggested that crossmodal integration arises prior to spatial selective attention. An important difference between Driver's (1996) study and the ones discussed above is the modality over which selective attention was measured (auditory vs. visual, respectively). These potential modality asymmetries will be addressed in this study.

In Experiment 1 participants were asked to detect or localise one out of four visually presented talking faces that matched a concurrent auditory speech stream. Participants received a pre-trial cue indicating two, three or the four locations among which the target (if present) could be found. Independency between RTs and the number of cued faces would indicate that audiovisual binding occurs before the deployment of (visual) spatial attention.

Experiment 1

Methods

Participants

We recruited 32 naive observers without prior history of hearing problems and normal or corrected vision. Fourteen participants took part in Experiment 1a (12 women, mean age 20.3 years) and 18 in Experiment 1b (16 women, mean age 20.7 years). Data from 2 additional participants were removed because they performed at chance in at least one condition.

Apparatus and materials

Stimuli consisted of digital video recordings (720x576 pixels; 25 fps) of a male speaker pronouncing 102 sentences in Spanish (selected from short written stories) at a rate of ~6.25 syllables·s^{-1} (Marín Gálvez 1994). The utterances were individually clipped in 4.8 s fragments including only the central part of each sentence, with onset and offset ramped (320 ms; 280 ms fade-in and out, respectively) in both audio and video channels. The search displays always contained four faces (6º x 8.5º degrees in size, shown on black background), each articulating a different sentence, shown above, below, left and right of a central fixation point (6.5º eccentricity). Such an eccentricity allows for a successful processing of visual speech information (Paré et al. 2003). In the target present trials (all localization and half detection task trials) only one of the four talking faces matched the auditory stream. The other three visually incongruent distractor faces were pseudorandomly chosen from the set of videotaped sentences (i.e., each distractor sentence was presented the same number of times throughout the

experiment). In target absent displays (used in the remaining half of detection task trials) none of the four faces in the display matched the auditory stream.

The video-clips were presented in a 22'' CRT monitor (100 Hz refresh rate). Sounds were presented at ~75 dB(A) –as measured from the participant's head– via two loudspeakers located on either side of the monitor. The experimental protocol was run on a PC using DMDX software (Forster and Forster 2003).

Procedure

Participants sat at a table in a dark, sound-attenuated room, with their head resting on a chinrest with the eyes 60 cm away from the monitor.

INSERT FIGURE 1 ABOUT HERE

Each trial (see **Fig. 1**) began with a central fixation (0.8º diameter) that remained in the centre of the screen. Participants were instructed to maintain their gaze at fixation during the trial. After fixation onset (500 ms), 2-4 ovals (6º x 8.5º visual angle) cued the potential target locations for 800 ms. This manipulation ensured identical sensory density across different display sizes (Louise et al. 2007). The pre-cued locations were randomly selected from trial to trial and equiprobable (within each set size) throughout the experiment. The cue was followed, after 1000 ms, by the search array including four talking faces plus a central auditory speech stream. In both the detection and localization tasks participants were instructed to press a key ("0" in the numeric keypad) as soon as they had come to a decision, but without sacrificing accuracy (RTs were calculated to this response). Participants were then prompted to provide a localization or detection response without time pressure (depending on the task).

Experiment 1a (detection task). Participants judged the presence/absence of a face matching the auditory sentence by keypress (response mapping counterbalanced across participants). In the target present trials (50%) the target could appear in either of the pre-cued locations equiprobably. Trials were presented in three blocks (of 144 trials; block order counterbalanced) by display size (i.e., number of pre-cued locations), following Fujisaki et al. (2006). Short resting breaks were introduced every 18 trials. Each block was preceded by 20 training trials not included in the analyses.

Experiment 1b (localization task): Participants were asked to locate the face matching the auditory spoken message using arrow keys (up, down, right or left arrows). Targets could appear in either of the pre-cued locations equiprobably. Subjects performed 9 blocks (three blocks of 24 trials for each display size) pseudo-randomly interleaved. Participants received 10 training trials prior to the first block of each condition.

Data analyses

Correct RT data were log transformed (Howell 2008; see Tables 1 and 2 for the mean Linear RTs)[2] and band-pass filtered ±2SDs for each participant and condition. RTs and accuracy data were regressed (linear) against display size (the number of cued faces) for each participant. The average slope and intercept of the regression lines were used for analyses. Two predictions derive from the hypothesis that audiovisual speech binding proceeds efficiently: First, regression slopes should be near zero (Experiment 1a and 1b), and second, the slope of target present trials should be similar to the slope of the

[2] The pattern of results remained the same when analysing the non-transformed data. Furthermore, we analysed all the experiments using RT as the dependent measure and display size as independent. Main effects and interactions mirrored the slope effects. For the sake of simplicity and clarity, we only report the analyses of search slopes throughout the article.

target absent trials (Experiment 1a only). Just like in previous studies using dynamic multisensory scenarios (Fujisaki et al. 2006; van der Burg et al. 2008 a), we expected long RTs, given that a certain amount evidence supporting the matching (or non-matching) nature of a time-varying event must be accumulated before response. Accuracy data were analyzed according to the nature of response set in each task. In the detection task (Experiment 1a), hits and false alarm rates were used to compute d′ for every participant and condition; Macmillan and Creelman 1991). In the localization task (Experiment 1b), to compensate the differences in chance level across display size (2, 3, or 4 alternatives) we used [accuracy=hit rate-(error rate/(number of possible responses - 1))] (Fig.2).

INSERT FIGURE 2 ABOUT HERE

Results

Experiment 1a: Detection task

The average slope for target present (.05 ± .03; Log(ms)/item ± SD) and target absent (.07 ± .05) trials indicated a significant increase in search times increased with set-size ($t(13)=6.5$, $p<.001$; $t(13)=5.8$, $p<.001$, respectively). The different intercepts between target present and absent conditions ($t(13)=4.56$, $p=.001$) indicated slower RTs for target absent (3.27 ± .18) than for target present trials (3.18 ± .16).

Target-absent trials had a steeper slope than target present trials (1.63 times steeper, $t(13)=2.93$, $p<.05$). This difference is short of the 2:1 ratio abiding to a strict self-terminating serial model, whereby observers need to scan over all the visual elements of the array to arrive to a conclusion in the target absent condition whereas, on average, only half the total number of items must be inspected in the target present

condition (Treisman and Gelade 1980)[3]. However, the small set sizes used here makes the slopes more amenable to deviations from this rule. Indeed, a variety of decisional (i.e., criterion) and/or processing strategies (i.e., participants rejecting more than one item at a time) can produce ratios smaller than 2:1 (see Wolfe 2003).

The *d'* data indicated a significant decline in accuracy as display size increased (*d'*=2.8, 2.3 and 2.2 for 2, 3 and 4 faces respectively; slope= -.28 ± .38, t(13)= -2.71, p=.02; intercept= 3.3 ± 1.4). This suggests that effects observed in the RT data did not derive from speed–accuracy tradeoffs.

Experiment1b: Localization task

RTs increased with set-size (slope= .07 ± .005, t(17)=12.43, p<.001, relative to 0; intercept= 3.22 ± .11). Although there was a trend toward a performance decrement with set-size (93%, 92% and 90% for 2, 3 and 4 faces, respectively) slope did not differ significantly from zero (slope= -.014 ± .04, t(17)=-1.5, p=.135; intercept= .95 ± .10).

Discussion

The results of Experiment 1 support and extend previous findings suggesting that visual selective attention is required prior to perceptual binding of correlated audiovisual objects (Andersen et al. 2008; Fairhall and Macaluso 2009; Fujisaki et al. 2006; Talsma and Woldorff 2007; van de Par and Kohlrausch 2004), and show that the pairing audiovisual events amongst visual distractors must proceed serially, even in the case of ecological and over learned stimuli such as speech. Indeed, these results appear to be at odds with the idea that audiovisual integration occurs prior to, and can determine,

[3] While non-zero slopes are generally taken to indicate that processing of the array elements depends on limited capacity resources, contrasting theories have explained set-size effects in terms of noisy, parallel processing (e.g., Palmer et al., 2000).

spatial selective attention, as suggested by van der Burg et al. (2008a, b) and Driver (1996). The discrepancy between Driver's findings and the present results in particular is remarkable because both studies involved audiovisual speech. One important difference is that, in contrast with Driver's *auditory* selective attention task, the present experiment examined the role of sound on *visual* selective attention. One of the key differences between auditory and visual processing is, precisely, the radically different encoding of space (e.g., Justus and List 2005; Kubovy 1988; Kubovy and Van Valkenburg 1995). For this reason, we examined whether similar constraints are to be found when search involves selective attention in the auditory domain.

INSERT FIGURE 3 ABOUT HERE

Experiment 2

Method

We tested 20 participants in the detection task (Experiment 2a; 5 women, 21.7 years old) and 18 in the localization task (Experiment 2b; 14 women, 21 years old). New audiovisual search displays were created based on the materials and procedure of Experiment 1. Four different spoken messages were presented simultaneously from different loudspeakers (using a 5.1 Dolby system) arranged in a semi-circle of radius 120 cm centred on the participant's head in the horizontal plane at angular positions −60° and −20° 20° and 60° (see **Fig. 3**). This distance allows for the separability (intelligibility and localization) of speech in multitalker displays (Brungart & Simpson 2005; Eramudugolla et al. 2008; Ericson et al. 2004, Simpson et al. 2006). Two, three,

or all four spoken messages were cued (using a schematic layout of the loudspeakers on the screen), and participants were asked to detect or to localize the one matching the talking face. Participants sat 1.80m away from the monitor used to display the (single) talker face (4.11º x 6.17 º, black background).

INSERT FIGURE 4 ABOUT HERE

Results

Experiment 2a: Detection task

Search slopes were not different from zero in the detection task (target present = -.004 ± .04; Log (ms)/item ± SD; t(19)=-.48, p=.634; target absent = .011 ± .034; t(19)= 1.44, p=.16) although target present and target absent slopes differed from one another due to their opposing numerical trends; t(19)=2.6, p=.02). Intercepts revealed faster responses in the target present than in the target absent condition (3.39 ± .09 vs. 3.54 ± .89; t(19)=7.09, p<.001). Thus, performance in the search task was difficult overall, but mostly unaffected by the number of items in the search set, a pattern that suggests an efficient, attention-free search. The accuracy analyses confirmed this, as the average slope of d' was not significantly different from zero (-0.15 ± 0.60, t(19)=-1.25, p=.22).

Experiment 2b: Localization task

Contrary to the detection results, localization performance declined significantly with set-size (RT slope= .04 ± .02, t(17)= 9.18, p<.001; intercept= 3.36 ± .08; d' slope= -.04

± .04, t(17)=10.75, p=.001; intercept= .92 ± .07). This suggests an effortful search and rules out speed–accuracy tradeoffs.

Discussion

Detection of audiovisual correlation amidst multiple competing speech streams is not contingent on the number of relevant utterances, suggesting efficient search. This is in stark contrast with localization that slowed down as the number of competing auditory streams increased, a pattern consistent with an attentive, effortful search process[4]. A possible account for the poorer localization performance in the auditory search task is that the matching auditory stream could have been mislocated towards the position of the central articulating face (by way of the ventriloquist illusion; Bertelson et al. 1994; Driver 1996; Spence and Driver 2000), creating spatial confusion with the distractors. Yet, a closer inspection of the data does not support this explanation because localization was overall accurate (i.e., participants extracted sufficient spatial information about the target; see **Fig. 4**) and target mislocalizations toward the matching speech signal or away from it were just as likely[5].

General Discussion

[4] Furthermore, this pattern of results generalized through different spatial layouts of the loudspeakers. That is, in a new control experiment the loudspeakers were arranged at the angles of an imaginary square centered on participants head (at 45º above and below the left and right ears, so that each loudspeaker was at an equal distance from the screen, 110 cm). Participants sat at approximately 80 cm from the screen. We found, again, that in the localization task (N=18) RT increased linearly with the number of attended elements (slope= .048 ± .035; intercept= 3.32 ± .13; t(17)=5.81, p<.001), whereas in the detection task (N=18) the mean search time was roughly constant regardless of the number of auditory distracters in the display, both when the target was present (slope= .011 ± .06, intercept= 3.41 ± .18, t(17)=-.45, p=.625) and when it was absent (slope= -.01 ± .06, intercept= 3.56 ± .18 ; t(17)=.754, p=.46).

[5] Analyses performed on the four cued condition of the localization task indicated that the proportion of trials in which participants incorrectly misallocated the targets –when presented in the extreme positions- towards the more central loudspeakers (i.e., closer to the location of the speaking face) was low (6%), and similar to the proportion of trials in which participants misallocated centrally presented targets towards the extremes (4%). This clearly suggests that the set-size effects observed in the localization task was not a consequence of the ventriloquism effect.

We examined whether the extraction of audiovisual correspondence requires spatial selective attention. Experiment 1 revealed that finding audiovisual correspondence amongst speaking faces is mediated by attentive search; That is, the efficiency to detect and locate audiovisual coincidence fell off as the number of talking faces increased, a trend that was steeper in target absent displays than in target present ones (detection task). The implication is that audiovisual correlation fails pop-out, as a pre-attentive multisensory binding hypothesis would predict (e.g., Van der Burg et al. 2008 a, b). The present results seem to be at odds with previous findings showing that the ventriloquist effect does not depend on visual attention (Bertelson et al. 2000a; Vroomen et al. 2001) or awareness (Bertelson et al. 2000b).

On the other hand, our findings are in line with evidence showing attention modulation of crossmodal integration (Alsius et al. 2005 2007; Andersen et al. 2008; Fairhall and Macaluso 2009; Fujisaki et al. 2006; Mozolic et al. 2008; Talsma et al. 2007; Tiippana et al. 2004; Tuomainen et al. 2005), and support the claim that attention mediates binding of object features that differentiate the target from distractors (i.e., auditory and visual correlated speech, in the present case; Treisman and Gelade 1980; Cinel et al. 2002). Unlike in Alsius et al. 2005, however, in this case the limitation seems to result mainly from unisensory capacity limitations within the visual system. This is because, auditory search (Experiment 2) resulted in automatic detection and visual search (Experiment 1) resulted in serial processing, when searching for exactly the same event (a speaking face). Thus, in Experiment 1, it is likely that the relevant (dynamic) visual information in the talking faces had to be extracted individually for each item in the visual scene, before the binding process could take place.

The result of Experiment 1 is surprising given previous studies where spatially non-informative but synchronized auditory events dramatically improve visual search performance (Iordanescu et al. 2007; Van der Burg et al. 2008a, b). The interpretation of these prior findings is that the sound enhanced the saliency of the matching visual event (with respect to the surrounding items), thereby releasing the selection system from the limitations imposed by visual attention. In our results, instead, audiovisual coincidence in speech did not release the visual attention system from the capacity limitations leading to inefficient search amongst faces. As hypothesized in the introduction, this discrepancy (Iordanescu et al. and Van der Burg et al. on the one hand, and Fujiaki et al. and the present results on the other) could be rooted in the informational complexity of the stimuli to be matched. In our experiments the time varying auditory stream did not provide with sufficiently salient spectral-temporal landmarks to link with the corresponding variations of the visual target. Instead, in van der Burg et al's experiments the accessory auditory stimulus was a salient singleton that could be quickly encoded for match against the visual items. One can thus reconcile the different outcomes within a framework based on perceptual load within the sensory modalities involved in crossmodal binding, as recently proposed by Talsma et al. (2010; see also Van der Burg et al., 2010). We argue that visual spatial selective attention is a crucial limiting factor in audiovisual binding of complex dynamic stimuli. Enhancement of visual selection based on fast pre-attentive binding will result only if the acoustic stimulus provides a distinctive (temporal) anchor.

Experiment 2 revealed that audiovisual correspondence in speech can be efficiently extracted amongst auditory distractors. These data fits well with previous findings by Driver (1996) suggesting that, provided visual attention is already allocated to the corresponding visual stream, crossmodal pairing can occur prior to the

deployment of auditory attention. That is, the sight of articulatory gestures may enhance processing of the corresponding auditory features and benefit search. These data are also consistent with object-based accounts of selective attention in audiovisual contexts, whereby attention to a unimodal component of a multisensory object spreads to other modalities, even when these are spatially misaligned (Busse et al. 2005; Fiebelkorn et al. 2010). Similarly, in the present experiment the auditory stream, albeit spatially misaligned, may be pulled into the attentional spotlight thanks to the correlated visual stimulation.

Detecting temporally correlated audiovisual speech needs attentive search when it is embedded amongst similar visual distractors, but it is efficient when the competition occurs amongst auditory distractors. Could differences in display homogeneity between targets and distractors be responsible for the reversed outcomes of visual and auditory search experiments (e.g., Duncan and Humphreys 1989)? We do not favor this interpretation because RTs were overall faster in the visual (less efficient) than in the auditory (more efficient) search tasks. We argue instead that the present dissociation points to a fundamental difference between auditory and visual spatial attention as a limiting factor for extraction of audiovisual correspondence, in particular, the obligatory encoding of spatial location in vision, but not in audition (Kubovy 1988; Kubovy and Van Valkenburg 2001). While both the auditory and visual modalities can represent stimulus localization in distal (extrapersonal) space, in vision, access to a feature value of an object (i.e., color) involves the obligatory encoding of its spatial location (e.g., Huang and Pashler 2007; Huang, et al. 2007). In audition, however, one can access the content of an auditory event without a sense about its spatial origin (e.g., Julesz and Hirsh 1972).

We propose that the present pattern of results can be explained as follows: Audiovisual matching of speech may occur automatically as long as access to the spatial location of the event is not required. Because encoding spatial location is obligatory in vision, when audiovisual matching requires visual selection it will be necessarily limited also by spatial selection, and thus will not occur automatically. Instead, given that encoding of spatial location is not mandatory for audition, when audio-visual matching requires auditory selection then search can proceed effectively as long as the task does not imply retrieving spatial location of the event. Audiovisual matching based on auditory selection will however abide to an attentionally demanding mode if spatial localization is required.

Acknowledgments

This work was supported by grants PSI2010-15426 and Consolider INGENIO CSD2007-00012 (MICINN), Generalitat de Catalunya (SRG2009-092), and European Research Council (StG-2010 263145).

References

Alsius A, Navarra J, Campbell R, Soto-Faraco S (2005) Audiovisual integration of speech falters under high attention demands. Curr. Biol. 15: 839-843.

Alsius A, Navarra J, Soto-Faraco S (2007) Attention to touch weakens audiovisual speech integration. Exp Br Res 183: 399-404.

Andersen TS, Tiippana K, Laarni J, Kojo I, Sams M (2008) The role of visual spatial attention in audiovisual speech perception. Speech Comm. 51: 184-193.

Bertelson P, de Gelder B (2004) The psychology of multimodal perception In: C. Spence and J. Driver (Eds), Crossmodal space and crossmodal attention. Oxford University Press, Oxford, pp. 141-177.

Bertelson P, Pavani F, Ladavas E, Vroomen J, de Gelder B (2000a) Ventriloquism in patients with unilateral visual neglect. Neuropsychologia 38: 1634-1642.

Bertelson P, Vroomen J, de Gelder B, Driver J (2000b) The ventriloquist effect does not depend on the direction of deliberate visual attention. Percept. Psychophys. 62: 321-332.

Bertelson P, Vroomen J, Wiegeraad G, de Gelder B (1994) Exploring the relation between McGurk interference and ventriloquism. Proceedings of the International Congress on Spoken Language Processing. Yokohama, Japan, pp. 559-562.

Bichot NP, Rossi AF, Desimone R (2005) Parallel and serial neural mechanisms for visual search in macaque area V4. Science 308: 529-534.

Brungart DS, Simpson BD (2005) Improving Multitalker Speech Communication with Advanced Audio Displays. Paper presented at the New Directions for Improving Audio Effectiveness- Meeting Proceedings RTO-MP-HFM-123, Neuilly-sur-Seine, France.

Busse L, Roberts KC, Crist RE, Weissman DH, Woldorff MG (2005) The spread of attention across modalities and space in a multisensory object. PNAS 102: 18751-18756.

Cinel C, Humphrey GW, Poli R (2002) Cross-modal illusory conjunctions between vision and touch. J. Exp. Psychol. Hum. Percept. Perform. 28: 1243-1266.

Colin C, Radeau M, Soquet A, Demolin D, Colin F, Deltenre P (2002) Mismatch negativity evoked by the McGurk-MacDonald effect: a phonetic representation within short-term memory. Clin. Neurophysiol. 113: 495-506.

Driver J (1996) Enhancement of selective listening by illusory mislocation of speech sounds due to lip-reading. Nature 2: 66-68.

Duncan J, Humphreys G (1989) Visual search and stimulus similarity. Psycho. Rev 96: 433-458.

Eramudugolla R, McAnally KI, Martin RL, Irvine D, Mattingley JB (2008) The role of spatial location in auditory search. Hearing Res 238: 139-146.

Ericson MA, Brungart,DS, Simpson BD (2004) Factors that influence intelligibility in multitalker speech displays. Int. J. Aviat. Psychol. 14: 311-332.

Fairhall SL, Macaluso E (2009) Spatial attention can modulate audiovisual integration at multiple cortical and subcortical sites. Eur. J. Neurosci. 29: 1-11.

Fiebelkorn IC, Foxe JJ, Molholm S (2010) Dual Mechanisms for the Cross-Sensory Spread of Attention: How Much Do Learned Associations Matter? Cereb Cortex, 20: 109-120.

Forster KI, Forster JC (2003) DMDX: A windows display program with millisecond accuracy. Behavior Research Methods, Instruments, & Computers 35: 116-124.

Fujisaki W, Koene A, Arnold D, Johnston A, Nishida S (2006) Visual search for a target changing in synchrony with an auditory signal. P. Roy. Soc. B, 273: 865–874.

Howell DC (2008) Fundamental statistics for the behavioral sciences (6th Edition). Belmont, CA: Duxbury Press. (1st Edition 1985).

Huang L, Pashler H (2007) A boolean map theory of visual attention. Psychol. Rev. 114: 599-631.

Huang L, Tresiman A, Pashler H (2007) Characterizing the limits of human visual awareness. Science 317: 823-825.

Iordanescu L, Guzman-Martinez E, Grabowecky M, Suzuki S (2008) Characteristic sounds facilitate visual search. Psychon. B. Rev. 15: 548-554.

Joseph JS, Chun MM, Nakayama K (1997) Attentional requirements in a ‘‘preattentive’’ feature search task. Nature 387: 805-807.

Julesz B, Hirsh IJ (1972) Visual and auditory perception – an essay of comparison in human communication. In David, E.E and Denes, P (Eds), Human communication: A unified view. McGraw-Hill, New York, pp. 283-335.

Justus T, List A (2005) Auditory attention to frequency and time: an analogy to visual local-global stimuli. Cognition 98: 31-51.

Kamachi M, Hill H, Lander K, Vatikiotis-Bateson E (2003) “Putting the Face to the Voice”: Matching Identity across Modality. Curr Biol 13: 1709-1714.

Kubovy M (1988) Should we resist to the seductiveness of the space: time::vision: audition analogy? J. Exp. Psychol. Human 14: 318-320.

Kubovy M, Van Valkenburg J (1995) Auditory and visual objects. Cognition 80: 97-126.

Lavie N (2005) Distracted and confused?: selective attention under load. TICS, 9(2), 75-82.

Louie EG, Bressler DW, Whitney D (2007) Holistic crowding: Selective interference between configural representations of faces in crowded scenes. J. Vis. **7**: 1-11.

MacMillan NA, Creelman CD (1991) Detection theory: A user's guide. Cambridge: Cambridge University Press.

Marín Gálvez R (1994) La duración vocálica en español. Estudios de Lingüística de la Universidad de Alicante 10: 213-226.

McElree B, Carrasco M (1999) The temporal dynamics of visual search: Evidence for parallel processing in feature and conjunction searches. J. Exp. Psychol. Human. 25: 1517-1539.

McGurk H, MacDonald J (1976) Hearing lips and seeing voices. Nature 265: 746-748.

Mozolic JL, Hugenschmidt CE, Peiffer AM, Laurienti PJ (2008) Modality-specific selective attention attenuates multisensory integration. Exp. Brain Res. 184: 39-52.

Navarra J, Alsius A, Soto-Faraco S, Spence C (2009) Assessing the role of attention in the audiovisual integration of speech. Inform. Fusion 11: 4-11.

Palmer J, Verghese P, Pavel M (2000) The psychophysics of visual search. Vision Res. 40: 1227-1268.

Paré M, Richler R, ten Hove M, Munhall KG (2003) Gaze Behavior in Audiovisual Speech Perception: The Influence of Ocular Fixations on the McGurk Effect. Percept. Psycho. 65: 553-567.

Pickering MJ, Garrod S (2006) Do people use language production to make predictions during comprehension? TICS, 11: 105-110.

Simpson B, Brungart D, Iyer N, Gilkey R, Hamil J (2006) Detection and localization of speech in the presence of competing speech signals. Proceedings of the 12th International Conference on Auditory Display, London, UK June 20-23.

Soto-Faraco S, Navarra J, Alsius A (2004) Assessing automaticity in audiovisual speech integration: evidence from the speeded classification task. Cognition 92: B13-B23.

Spence C, Driver J (2000) Attracting attention to the illusory location of a sound: reflexive crossmodal orienting and ventriloquism. Neuroreport 26: 2057-2061.

Spence C, Santangelo (2009) Capturing spatial attention with multisensory cues: a review. Hearing Res. http://dx.doi.org/10.1016/j.heares.2009.04.015

Talsma D, Doty TJ, Woldorff MG (2007) Selective attention and audiovisual integration: Is attending to both modalities a prerequisite for early integration? Cereb. Cortex 17: 691-701.

Talsma D, Senkowski D, Soto-Faraco S, Woldorff MG (2010) The multifaceted interplay between attention and multisensory integration. Trends Cogn Sci, 14: 400-10.

Tiippana K, Andersen TS, Sams M (2004) Visual attention modulates audiovisual speech perception. Eur. J. Cog. Psychol. 16: 457-472.

Treisman A, Gelade G (1980) A feature integration theory of attention. Cognitive Psychol., 12: 97-136.

Tuomainen J, Andersen TS, Tiippana K, Sams M (2005) Audio-visual speech perception is special. Cognition, 96: B13-B22.

Van der Burg E, Olivers CNL, Bronkhorst AW, Theeuwes J (2008a) Pip and pop: Non-spatial auditory signals improve spatial visual search. J. Exp. Psychol. Human. 34: 1053-1065.

Van der Burg E, Olivers CNL, Bronkhorst AW, Theeuwes J (2008b) Audiovisual events capture attention: Evidence from temporal order judgements. J Vision 8: 1-10.

Van der Burg E, Cass J, Olivers CN, Theeuwes J, Alais D (2010) Efficient visual search from synchronized auditory signals requires transient audiovisual events. PLoS One, 5(5), e10664.

van de Par S, Kohlrausch A (2004) Visual and auditory object selection based on temporal correlations between auditory and visual cues. In Proceedings of the 18th International Congress on Acoustics. Kyoto, Japan, pp.2055-2058.

Van Wassenhove V, Grant KW, Poeppel D (2005) Visual speech speeds up the neural processing of auditory speech. PNAS 102: 1181-1186.

Vroomen J, Bertelson P, de Gelder B (2001) Auditory-visual spatial interactions: automatic versus intentional components. In B., de Gelder, E. de Haan, C. Heywood (Eds.), *Out of Mind* . Oxford University Press, Oxford, pp. 140-150.

Welch RB (1999) Meaning, attention, and the "unity assumption" in the intersensory bias of spatial and temporal perceptions. In G. Ashersleben, T. Bachmann, J. Müsseler (Eds.), Cognitive contributions to the perception of spatial and temporal events. Elsevier, Amsterdam, pp. 371-387.

Welch RB, Warren DH (1980) Immediate perceptual response to intersensory discrepancy. Psychol. Bull. 3: 638-667.

Wolfe JM (2003) Moving towards solutions to some enduring controversies in visual search. Trends Cog. Sci. 7: 70-76.

Figure captions

Figure 1. On each trial of Experiment 1, participants were shown a fixation point, followed by a spatial cue and then a search array, which consisted of four talking faces, each one articulating a different utterance, while hearing a single auditory speech stream. Participants had to either locate the face matching the heard sentence, or to perform a detection task (i.e., decide whether one of the faces in the display matched or not the spoken utterance).

Figure 2. Log(RT)s (upper panels) and accuracy (lower panels) for detection (left panel) and localization task (right panel) in Experiment 1a and 1b (respectively). Error bars indicate the standard error of the mean. The symbols represent the observed results, and the lines are the fits of the data obtained by linear regression analysis.

Figure 3. Layout of the experimental set up used in Experiment 2. The auditory utterances were delivered through external loudspeakers at the depicted locations. The observers had their head aligned with the centre of the screen and their gaze directed at the fixation point.

Figure 4. Log(RT)s (upper panels) and accuracy (lower panels) for detection (left panel) and localization task (right panel) in Experiment 2a and 2b (respectively). Error bars indicate the standard error of the mean. The symbols represent the observed results, and the lines are the fits of the data obtained by linear regression analysis.

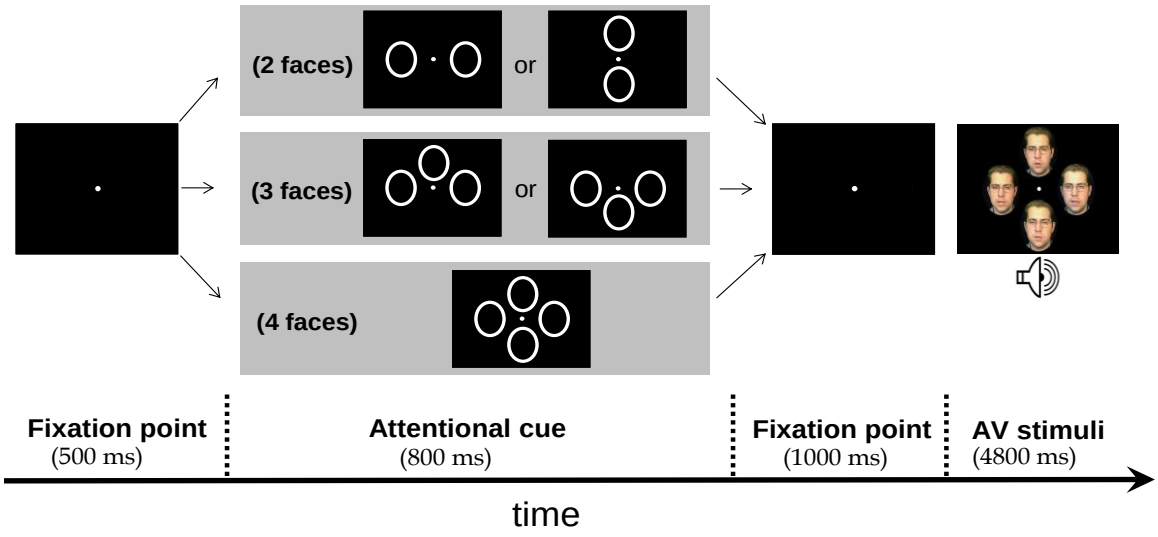

Figure 1. ALSIUS and SOTO-FARACO

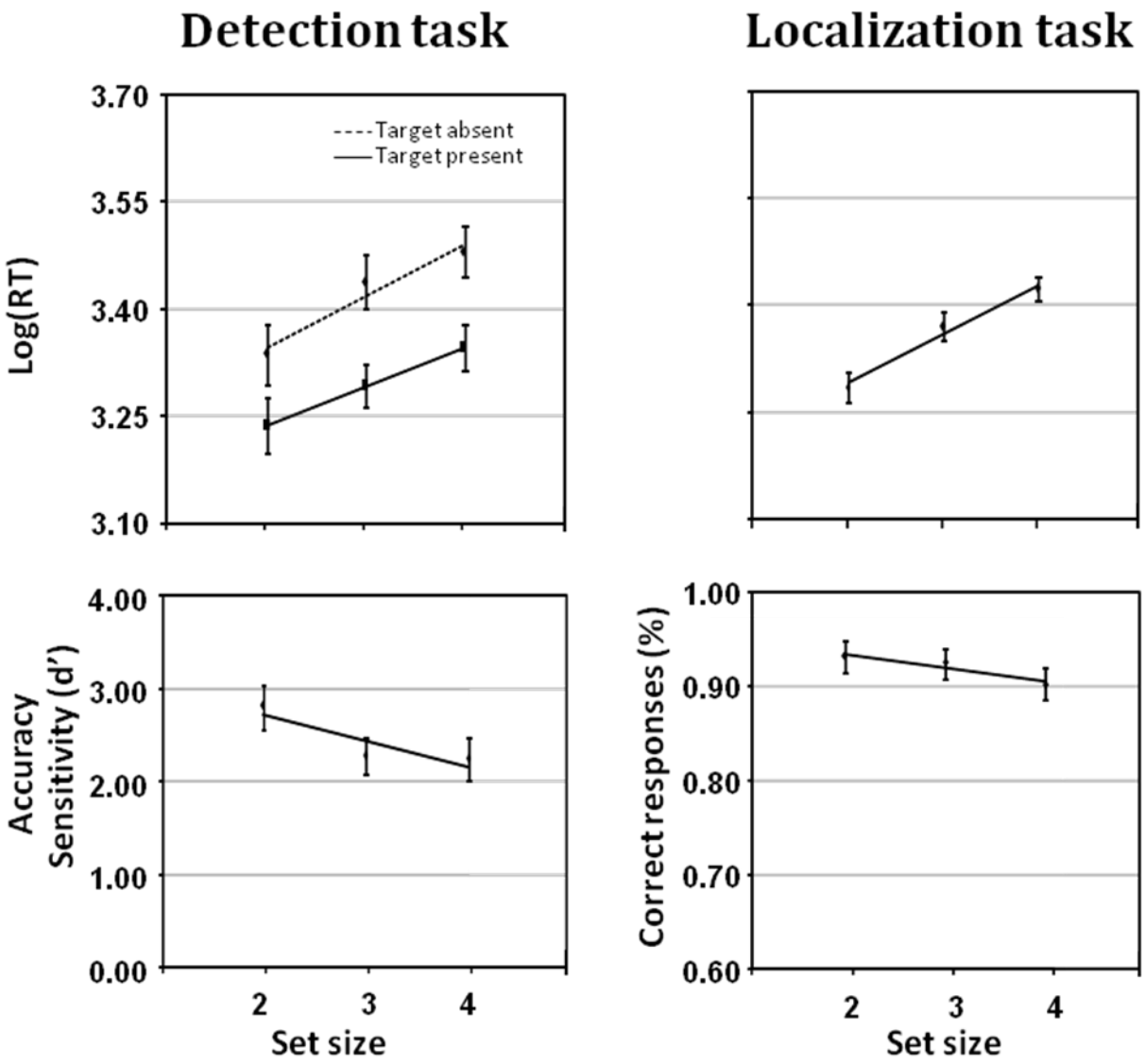

Figure 2. ALSIUS and SOTO-FARACO

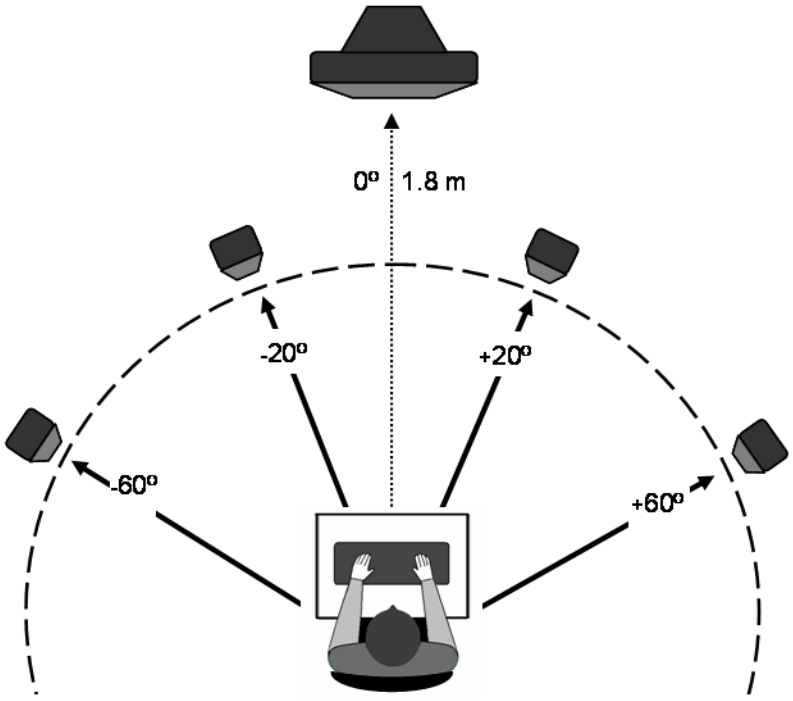

Figure 3. ALSIUS and SOTO-FARACO

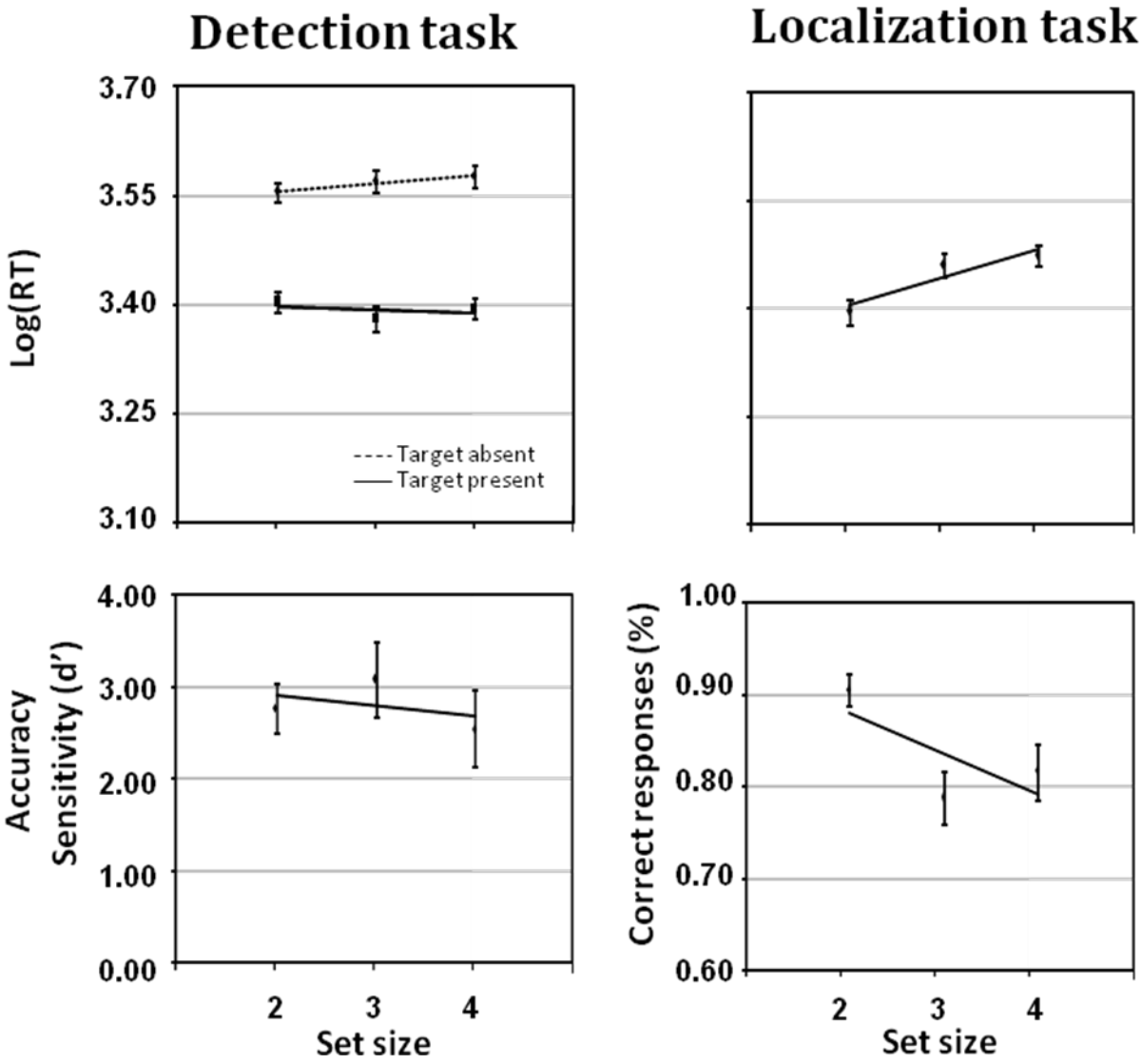

Figure 4. ALSIUS and SOTO-FARACO

Additional tables

		LOCALIZATION TASK				
		2	**3**	**4**	***slope***	***Intercept***
EXP1b	Linear RT	2004	2447	2716	355	1678
	Log (RT)	3.28	3.37	3.42	0.07	3.22
EXP2b	Linear RT	2580	3000	3104	262	2370
	Log (RT)	3.39	3.46	3.47	0.04	3.36

Mean correct Linear RT and Log (RTs) for each Condition and Experiment in which participants performed a localization task.

		DETECTION TASK									
		Target Present					Target Absent				
		2	3	4	*slope*	*intercept*	2	3	4	*slope*	*Intercept*
EXP1a	Linear RT	1855	2078	2340	242	1605	2343	2924	3195	425	1969
	Log (RT)	3.24	3.29	3.34	0.05	3.18	3.33	3.43	3.48	0.07	3.27
EXP2a	Linear RT	3689	2501	2567	-60.72	2708	3628	3791	3916	143	3490
	Log (RT)	3.42	3.39	3.41	-0.00	3.39	3.56	3.57	3.59	0.01	3.54

Mean correct Linear RT and Log (RTs) for each Condition and Experiment in which participants performed a Detection task.