Benefits of Multisensory Cues in a Realistic Target Discrimination Task

Caterina Cinel[1], Jacobo Fernandez Vargas[1], Luca Citi[1] and Riccardo Poli[1]

Abstract—**This study is part of large project aiming at developing brain-computer interfaces capable of enhancing decision-making performance in multisensory environments. In this paper, we present results from an experiment exploring the possibility of increasing accuracy of users engaged in the discrimination of difficult realistic targets, in the presence of multisensory cues. More specifically, in our experiment users are asked to discriminate between targets rapidly appearing in a indoor environment. Auditory or audio-visual cues may precede the target, warning the users about the location of its imminent appearance. Results show that if the timing of the audio-visual cues is right, they can be more effective at preparing users to respond to the imminent target, compared to audio-only cues, thereby increasing user performance.**

I. Introduction

In situations where fast reactions to sudden stimuli are critical, spatial cues preceding such stimuli can help improve performance by triggering the shift of attention to the relevant location before a target stimulus appears. This has been demonstrated in many experimental settings, particularly in simple tasks using the classic spatial cueing paradigm [1]–[3]. Typical studies show that a cue at the location where a task-relevant stimulus is going to appear — called *exogenous cue* — will improve both response speed and accuracy, compared to when no cue is presented or to a cue presented at a different location (which can actually worsen performance as compared to no cue). A similar effect is observed with symbolic cues — called *endogenous cues* — such an arrow pointing towards the direction where the stimulus will appear, or a word (e.g., “left” or “right”), which provide spatial information without being in the cued location.

The effects of spatial cues on responses to events have also been examined in more realistic situations. For example, in driving simulations [4]–[6] or air traffic control simulations [7], [8] a spatial cue facilitates a prompter response to a potentially dangerous event, particularly in situations of high information load, which make situation awareness particularly challenging. However, most of what is currently known of how spatial cueing can affect performance, and when and how it can be more effective, is the result of investigations carried out within a *single modality* — the visual perception domain — and in relatively simple experimental conditions.

Research in multisensory perception has shown that *multisensory* cue-stimulus combinations can enhance performance [9]. This can have important implications in real-life tasks. For example, Begault [9] found that the performance in a visual search task of airline crew improved when using spatial auditory cues in the same location of the target, compared to when the cue was a warning message (“traffic, traffic!”) with no spatial information.

More recently Ho and Spence [5] conducted a series of experiments investigating the benefits of using auditory spatial cues in a simulated driving task. They examined the effects of nonspatial-nonpredictive, spatial non-predictive, and spatially-predictive sound cues, as well as symbolic predictive and spatially-presented symbolic predictive verbal cues. In all cases, auditory cues coming from the visual target relevant direction improved performance (see also other investigations on cross-modal links in attention, e.g., [4], [7]). Moreover, the advantage was enhanced when non-spatial, but semantically informative cues (the words “front” and “back”) were presented. Consistently with Ho and Spence [5], other studies [1], [5], [6], [10] have shown that spatial information cueing to the location or direction of a task-relevant visual stimulus can be provided not only by an exogenous cue but also by endogenous, meaningful directional cues — such as spoken words — presented in a non-relevant location (i.e., at the centre of the display, when the target-relevant visual stimulus is presented at left/right, top/bottom or front/rear locations).

As empirical evidence suggests that there is an advantage of providing cues and task-relevant stimuli in different sensory modalities, it is not clear whether providing multisensory cues might further enhance performance: for instance, would an audio-visual cue be more effective at shifting attention to the location of an imminent visual stimulus as compared to an audio-only cue? Our cognitive system seems to assume that multisensory events that are very close temporally or spatially belong together [11]–[13] and both behavioural and neural responses – including the shifting of attention — to the environment can be enhanced when the information provided is multisensory [14]–[17]. Furthermore, there are indications that multisensory cues might be more effective in some cases in attracting attention to a location as compared to unimodal cues [18], [19].

The aim of this study was twofold. Firstly, we wanted to investigate whether the effect of a spoken spatial cue, giving information about the approximate location of an imminent

The authors acknowledge support of the UK Defence Science and Technology Laboratory (Dstl) and Engineering and Physical Research Council (EPSRC) under grant EP/P009204/1. This is part of the collaboration between US DOD, UK MOD and UK EPSRC under the Multidisciplinary University Research Initiative.

[1]Caterina Cinel, Jacobo Fernandez Vargas, Luca Citi and Riccardo Poli are with the Brain Computer Interfaces and Neural Engineering Laboratory, School of Computer Science and Electronic Engineering, University of Essex, UK. Email: {ccinel,j.fernandez-vargas, lciti,rpoli}@essex.ac.uk.

task-relevant stimulus, is enhanced when presented in combination with a visual cue. Secondly, we investigated whether the effect of the spoken unimodal or multimodal cues were modulated by the interval between the cue and the target stimulus.

This study is part of a US-UK collaboration in a Multidisciplinary University Research Initiatives (MURI) program, that aims at developing brain-computer interfaces capable of enhancing decision-making performance in multisensory environments. Here we explore the possibility of increasing accuracy of users engaged in a difficult realistic discrimination task — the identification of suddenly appearing unidentified military personnel when patrolling/clearing an indoor environment — in the presence of additional informative verbal or audio-visual communication (e.g., from a control centre).

II. Testing multisensory cues in an indoor environment

To address the aims of our study, we performed an experiment where participants had to rapidly decide whether a character in uniform, appearing for a very short time on the left or right side of a poorly-lit realistic indoor scene, was wearing a cap or a helmet. In some of the experiment blocks, the target was preceded by a cue indicating whether the character was appearing on the left or the right. The cue could be either a spoken word ("left" or "right") only, or in in combination with a synthetic face uttering the cue word (with matching lip movements). Performance was then measured in terms of both accuracy and reaction times. More details are provided below.

A. Method

Participants were presented with displays showing a simplified image of a long corridor with doors on either side, and where a character wearing either a helmet or a cap was shown (in a posture suggesting the character is walking across the corridor), for 250 ms (see Fig. 3(a)). Participants had to decide, as rapidly as possible, whether the character was wearing a helmet or a cap (see Fig. 3(b)).

Each trial started with a display showing a fixation cross for 1 second. This was followed by a pre-stimulus display, where the corridor environment was shown and where a cue was presented or not. When a cue was presented, this was either an Auditory Cue (AC) or an Audio-Visual Cue (AVC). The AC consisted of a voice uttering either the word "left" or the word "right", to indicate the side of the corridor where the character would appear (with 100% validity) in the following display. The AVC consisted of a face uttering either the word "left" or the word "right". In trials where there was no cue, in the pre-stimulus displays either an empty corridor (No Cue, or NC) or a static silent face (Visual Cue, or VC) was presented. In AVC and VC displays, the face was shown at the centre top of the display, as shown in Fig. 3(c). The time interval between the pre-stimulus display and the target stimulus — also known as Stimulus Onset Asynchrony (SOA) — could be 500, 700 or 900 ms.

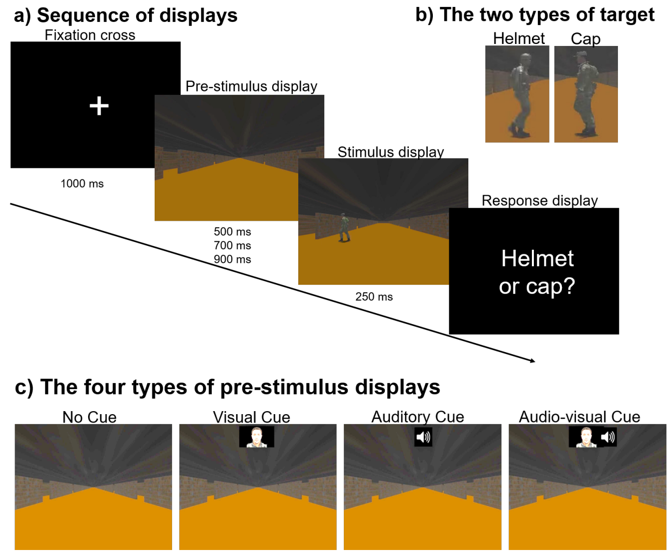

Fig. 1. a) The sequence of displays on each trial and their duration. The duration (SOA) of the pre-stimulus display varied randomly between 500, 700 and 900 ms, while the response display lasted until a response was given; b) The two types of target characters appearing in the stimulus display (helmet or cap); c) In each block, one of the four types of pre-stimulus displays was shown: No Cue, Visual Cue, Auditory Cue or Audio-Visual Cue.

The experiment consisted of 16 blocks of 48 trials each. In each block, the type of pre-stimulus display was fixed, with four blocks for each pre-stimulus display type. The block order was randomised for each participant, and each block was preceded by a display informing the participant about the type of pre-stimulus display of the block. At the end of each block, the mean accuracy for that block was displayed as a form of feedback. Reaction Times (RTs) and accuracy were recorded. Participants completed the experiment in a sound-proof lab room and sat about 90 cm from the computer screen. The experiment took about 45 minutes, including instructions, a short practice session and breaks between blocks.

B. Participants, Design and Ethical Approval

Thirty-three participants in total took part in the experiment. However, we discarded participants whose accuracy in the task was below 60% (three participants), and participants who did not complete the experiment (two participants).Therefore, here we present data from 28 participants in total. All participants had normal or corrected-to-normal vision and signed an informed consent form before taking part in the study.

The experiment had a 2-way within-subjects design. The two variables were the type of pre-stimulus display (Pre-Stimulus), with four levels (NC, VC, AC and AVC) and the SOA, with three levels (500, 700 and 900 ms). A 3$\times$4 within-subject ANOVA test was performed on RTs and accuracy.

The research received ethical approval by the UK Ministry of Defence (Application Number: 832/MoDREC/17) and the University of Essex in June 2017. All experiments were performed in accordance with the relevant guidelines and regulations.

III. Results

Reaction times longer than 1.5 second were considered outliers and discarded (on average 4% of trials were removed). In addition to this, as is typical in cognitive experiments, incorrect responses were also discarded from RT analyses.

A. Accuracy

The overall mean accuracy (across all participants and conditions) was 86.6% (SD = 5.4). However, accuracies varied depending on SOA and pre-stimulus type, as shown in Fig. 2. A 3×4 withing-subject ANOVA was performed. This showed that there was a significant main effect of Pre-Stimulus ($F(3,27) = 7.97$, $p < .001$) — represented by the grey dashed line in Fig. 2 — with AVC trials being the most accurate (88.6%, SD = 5.1), followed by AV trials (87.1%, SD = 5.7), NC trials (86.2%, SD = 6.9) and VC trials (84.6%, SD = 6.0). There was also a significant effect of SOAs ($F(2,27) = 4.43$, $p = .017$), with accuracy being lower with the 500-ms interval (85.6%, SD = 6.1), followed by the 700-ms (87%, SD = 5.2) and higher with the 900-ms interval (87.2%, SD =5.6). Finally, the interaction between the cue type and SOA was not significant ($F < 1$).

As there was a significant main effect of SOA, to look more specifically into it, we carried out three pairwise comparisons of the data from the three intervals, independently of pre-stimulus type. These showed a significant difference (Holm-Bonferroni adjusted) between 500-ms and 900-ms SOAs ($p = .033$), but not between 500-ms and 700-ms, nor between 700-ms and 900-ms SOAs (p-values $> .1$). The average accuracy values obtained with the three time intervals can been seen in Fig. 2, above the "Average" label.

We carried out a similar analysis to examine the main effect of Pre-Stimulus, independently of SOA (illustrated by the four points in the grey line labelled "Average" in Fig. 2): we tested the difference between all pairings of pre-stimulus type, with Holm-Bonferroni adjusted pairwise comparisons. These showed a significant difference between VC and AC trials ($p = .005$), and VC and AVC trials ($p > .001$). All other differences where non-significant (all p-values $> .1$), though the difference between NC and AVC trials approached significance ($p = .08$).

All in all, the data indicates that there was an advantage in increasing the SOA, with participants being more accurate with the 900-ms SOA. Furthermore, the type of pre-stimulus display also had an effect on accuracy. In particular, the highest accuracy was achieved in AVC trials, even compared to AC trials, suggesting that there is an advantage in providing spatial information in "multisensory form". The effect was enhanced by longer SOAs, indicating that more than 500 ms are needed to process the semantic information provided (as it was the case for AC and AVC trials) and to react to it (i.e., saccading and/or shifting and focusing the attention to the cued direction).

B. Reaction Times

The mean RT across all conditions was 489 ms (SD 107 ms). Mean RTs, according to SOA and pre-stimulus type, are displayed in Fig. 3. A 3×4 within-subjects ANOVA showed that there was a main effect of SOA ($F(2, 27) = 13.47$, $p < .001$), indicating that RTs were slower with shorter intervals, namely: 503 ms (SD = 109) for a 500-ms interval, 482 ms (SD = 109) for a 700-ms interval, and 492 ms (SD = 105) for a 900-ms interval (see Fig. 3). In particular, Holm-Bonferroni-adjusted pairwise comparisons of all SOA pairs (regardless of Pre-Stimulus) showed that RTs in the 500-ms interval were significantly slower than RTs in both the 700-ms interval and the 900-ms interval (both p-values $< .001$). The difference between the 700-ms and 900-ms intervals was non-significant ($p = .9$).

The main effect of Pre-Stimulus was not significant ($F(3, 27) = 2.04$, $p = .114$), but there was a significant interaction between Pre-Stimulus type and SOA ($F(6, 27) = 5.038$, $p < .001$). To further analyse and understand the interaction, we

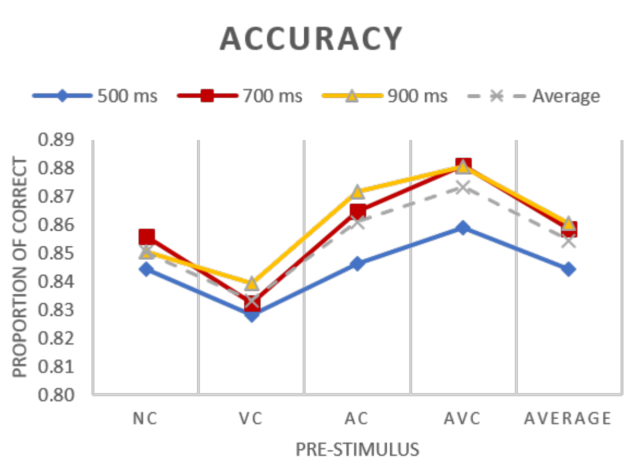

Fig. 2. Mean accuracies according to SOA (500, 700 and 900 ms) and pre-stimulus display type (NC, VC, AC and AVC).

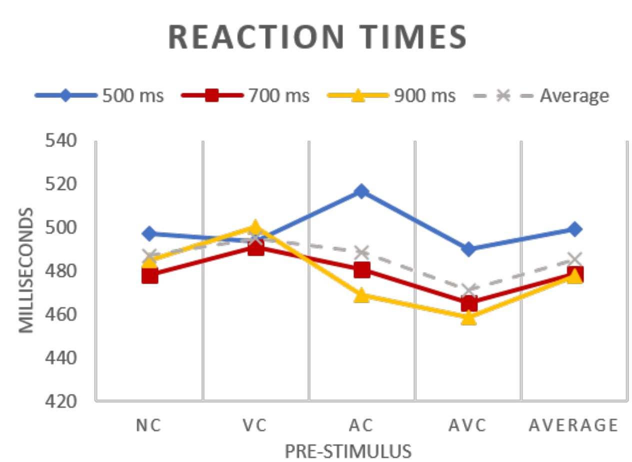

Fig. 3. Mean RTs according to SOA (500, 700 and 900 ms) and pre-stimulus display type (NC, VC, AC and AVC)

performed post-hoc analyses, where we analysed the effect of Pre-Stimulus within each SOA condition separately. The three ANOVAs showed that there was a main effect of Pre-Stimulus for 900-ms SOA trials ($F(3,27) = 4.94$, $p = .003$), but not for 500-ms ($F(3,27) = 1.84$, $p = 0.147$) or 700-ms trials ($F(3,27) = 1.72$, $p = .169$).

Overall, the results of our statistical analyses (for both accuracy and RTs) show that there is an advantage to providing spoken information about the location of the visual stimulus, provided that there is a long enough interval between cue and stimulus. Most importantly, our initial hypothesis that there might be an advantage resulting from the multisensory feature combination of the cue was confirmed (significant main effect of pre-stimulus type on accuracy).

The combined effect of SOA and pre-stimulus display on performance is illustrated in Fig. 4. This shows not only that the audio-visual cues provide an advantage over audio-only cues, but also that this effect is further enhanced by the longer SOAs. The criticality of the choice of timing is also illustrated by the AC at 500-ms condition, in which a dramatic drop in accuracy is caused by the presentation of the cue too close to the visual stimulus (at the 500-ms interval). VC trials were the least accurate and were amongst the slowest. This is not surprising, as the visual cue alone does not give any additional information, unlike the ACs and AVCs, and it might actually act as a distractor (for example, by diverting attention from the main stimulus location), compared to the NC conditions.

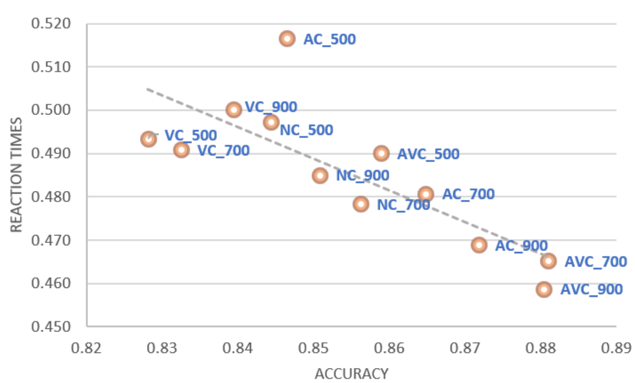

Fig. 4. Accuracy in relation to RTs for each of the twelve different conditions.

C. Discussion and Conclusions

The main purpose of this study was to investigate the effectiveness of auditory and audio-visual cues providing information about the location of an imminent task-relevant target, and whether this effect was mediated by the time interval between the cue and the target. Previous research suggests that multisensory cues can be more effective than unimodal cues [14]–[19]. However, it is no clear what the interval between cue and stimulus should be, particularly with *endogenous* spatial cues like those used in the current study. It is well known from the literature that valid *exogenous* cues improve performance when the stimulus is presented within about 300 ms from the cue, with intervals longer than that producing the opposite effect, known as *inhibition of return* (IOR; [20]). However, this does not necessarily happen with endogenous cues and Ho and Spence [5], [6] did not find any IOR when spoken words (”left” and ”right”) were used as endogenous cues, at SOAs of 200, 400 and 600 ms.

In our experiment, there were three different SOAs. We reasoned that the effect of audio and audio-visual cues might be accentuated with the longer intervals, as it would allow more time to process, and react to, the cue information. Indeed, we found that increasing the SOA improved performance particularly with AC and AVC trials. Most importantly, there was a clear benefit of presenting multisensory cues, compared to audio-only cues, which confirmed our initial hypothesis. In addition, there was a combined effect of cue-stimulus interval and type of cue, in which the effect of the multisensory cue increased with longer SOA, with the results that the best performance was obtained with the audio-visual cues presented 700 or 900 ms before the stimulus.

Our results have important implications for real-life circumstances in which cues are used with the purpose of increasing situation awareness, such as, for example, in flight control, during driving or patrol, in which users are engaged in difficult realistic discrimination tasks (e.g., the identification of suddenly appearing unidentified military personnel when patrolling/clearing an indoor environment) in the presence of additional informative verbal or audio-visual communication (e.g., from a control centre). They suggest that not only the right combination of cue modalities is critical, but also that the timing with which information is delivered is equally important.

In future research we will investigate the effects of modality and timing for other types of cues (e.g., non-spatial) and for even more realistic environments.

References

[1] B. Hommel, J. Pratt, L. Colzato, and R. Godijn, “Symbolic control of visual attention,” *Psychological Science*, vol. 12, no. 5, pp. 360–365, 2001.

[2] M. I. Posner, “Orienting of attention,” *Quarterly Journal of Experimental Psychology*, vol. 32, no. 1, pp. 3–25, 1980.

[3] M. I. Posner and Y. Cohen, “Components of visual orienting,” *Attention and performance X: Control of language processes*, vol. 32, pp. 531–556, 1984.

[4] C. Ho, H. Z. Tan, and C. Spence, “Using spatial vibrotactile cues to direct visual attention in driving scenes,” *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 8, no. 6, pp. 397–412, 2005.

[5] C. Ho and C. Spence, “Assessing the effectiveness of various auditory cues in capturing a driver’s visual attention.” *Journal of Experimental Psychology: Applied*, vol. 11, no. 3, p. 157, 2005.

[6] ——, “Verbal interface design: Do verbal directional cues automatically orient visual spatial attention?” *Computers in Human Behavior*, vol. 22, no. 4, pp. 733–748, 2006.

[7] S. Hameed, S. Jayaraman, M. Ballard, and N. Sarter, “Guiding visual attention by exploiting crossmodal spatial links: An application in air traffic control,” in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 51, no. 4. SAGE Publications Sage CA: Los Angeles, CA, 2007, pp. 220–224.

[8] M. K. Ngo, R. S. Pierce, and C. Spence, “Using multisensory cues to facilitate air traffic management,” *Human factors*, vol. 54, no. 6, pp. 1093–1103, 2012.

[9] D. R. Begault, “Head-up auditory displays for traffic collision avoidance system advisories: A preliminary investigation,” *Human Factors*, vol. 35, no. 4, pp. 707–717, 1993.

[10] A. R. Hunt and A. Kingstone, “Covert and overt voluntary attention: Linked or independent?” *Cognitive Brain Research*, vol. 18, no. 1, pp. 102–105, 2003.

[11] P. Bertelson and M. Radeau, “Cross-modal bias and perceptual fusion with auditory-visual spatial discordance,” *Perception & Psychophysics*, vol. 29, no. 6, pp. 578–584, 1981.

[12] L. Shams, Y. Kamitani, and S. Shimojo, “Illusions: What you see is what you hear,” *Nature*, vol. 408, no. 6814, p. 788, 2000.

[13] J. Vroomen and M. Keetels, “Perception of intersensory synchrony: A tutorial review,” *Attention, Perception, & Psychophysics*, vol. 72, no. 4, pp. 871–884, 2010.

[14] D. Alais, F. Newell, and P. Mamassian, “Multisensory processing in review: From physiology to behaviour,” *Seeing and Perceiving*, vol. 23, no. 1, pp. 3–38, 2010.

[15] G. A. Calvert, “Crossmodal processing in the human brain: Insights from functional neuroimaging studies,” *Cerebral Cortex*, vol. 11, no. 12, pp. 1110–1123, 2001.

[16] M.-H. Giard and F. Peronnet, “Auditory-visual integration during multimodal object recognition in humans: A behavioral and electrophysiological study,” *Journal of Cognitive Neuroscience*, vol. 11, no. 5, pp. 473–490, 1999.

[17] D. Talsma, D. Senkowski, S. Soto-Faraco, and M. G. Woldorff, “The multifaceted interplay between attention and multisensory integration,” *Trends in Cognitive Sciences*, vol. 14, no. 9, pp. 400–410, 2010.

[18] V. Santangelo and C. Spence, “Multisensory cues capture spatial attention regardless of perceptual load.” *Journal of Experimental Psychology: Human Perception and Performance*, vol. 33, no. 6, p. 1311, 2007.

[19] C. Spence and V. Santangelo, “Capturing spatial attention with multisensory cues: A review,” *Hearing Research*, vol. 258, no. 1-2, pp. 134–142, 2009.

[20] M. Mulckhuyse, D. Talsma, and J. Theeuwes, “Grabbing attention without knowing: Automatic capture of attention by subliminal spatial cues,” *Visual Cognition*, vol. 15, no. 7, pp. 779–788, 2007.