AURALLY AIDED VISUAL SEARCH IN DEPTH USING 'REAL' AND 'VIRTUAL' CROWDS

Jason S. Chan[1], Corrina Maguinness[1], Simon Dobbyn[2], Paul McDonald[3], Henry Rice[3], Carol O'Sullivan[2], Fiona N. Newell[1]
School of Psychology and Institute of Neuroscience[1], Department of Computer Science[2],
Department of Mechanical Engineering[3]
Trinity College Dublin
jason.chan@tcd.ie

Abstract

Auditory stimuli are known to improve visual target recognition and detection when both are presented in the same spatial location. However, these studies have focused on spatial congruency along the horizontal plane. To date, it is unknown whether the audio-visual spatial congruency is important for localisation in depth. In Experiment 1, we presented simple audio-visual stimuli presented in a congruent spatial location or the auditory stimulus was presented directly in-front or behind the visual stimulus. We find that participants are faster and more accurate when audio-visual stimuli are presented in the same location, compared to different depths. To increase the number of distracters and locations, a more complex scene was created and presented on a computer monitor. Virtual audio was recorded to be congruent or incongruent with the visual target. Participants were asked to locate a target agent (virtual person) amongst a varying number of distracters. Once again, we found participants were faster and more accurate when audio-visual stimuli were presented in the same location in the scene, compared to different "depths".

Introduction

It is well known that an auditory stimulus can aid the detection of the visual stimulus if they are presented in the same spatial location (Perrott, 1984; Perrott, Cisneros, McKinley, & D'Angelo, 1995). However, these stimuli only focused on the horizontal plane. Perrott and colleagues used a visual search paradigm while presenting an auditory stimulus that was either spatially congruent, incongruent, or absent. They found that the congruent auditory stimulus improved target detection, compared to the incongruent and sound absent conditions.

Multisensory depth perception is a little understood plane of existence. Some previous perceptual studies have inadvertently shown Temporal order effects when the visual and auditory stimuli were presented in a different plan in depth. Hirsh and Sherrick (1961) presented their visual stimuli on a computer screen while presenting the auditory stimuli through headphones and found participants were able to determine which stimulus was presented first. However, Zampini, Shore and Spence (2003) presented the auditory stimuli through loudspeakers located positioned at the same location as the visual stimuli found participants were significantly better able to determine the temporal order of the stimuli when the audio and visual stimuli were presented in different locations compared to the same spatial location. The conflicting results of these studies suggest that despite the small difference in depth between a visual and auditory stimulus, it can have profound effects on the perception of multisensory events.

Sugita and Suzuki (2003) explicitly looked at the role of depth in the perception of audiovisual temporal order judgements. They presented visual stimuli at different distances while the auditory stimuli were presented through headphones. They delayed the stimulus onset asynchrony (SOA) between the visual and auditory stimuli to simulate different depths. Participants find it difficult to determine whether an audio or visual stimulus was presented first when the stimuli are presented at a distance less than 20m (Lewald & Ehrenstein, 1998; Sugita & Suzuki, 2003). They argue that this is because of the temporal window of integration between vision and audition.

Previous multisensory studies have focused on the temporal interactions between vision and audition. In the current study, we explore the spatial interactions between these two modalities. We presented audio-visual stimuli from either the same or different depths. If congruent depth information is necessary for multisensory attention then we would expect incongruent audio-visual stimuli to impair accuracy and/or reaction time performance compared to the congruent audio-visual condition. However, if multisensory attention is not sensitive to depth information then we expect no difference in performance between the two conditions.

Experiment 1

Methods

Participants

Twenty-three (19 = female) participants between the ages of 17 years and 36 years (mean age = 24 years) took part in this experiment. All participants were right handed except for one. Participants had normal or corrected to normal vision and they did not report any hearing impairments.

Apparatus and Materials

The entire apparatus consisted of two semicircular arcs, located at 60cm and 120cm, directly in front of the participant. The location of each loudspeaker within each arc was -67.5°, -22.5°, 22.5°, 67.5° relative to the participant. In other words, each loudspeaker in the 120cm arc was directly behind its relative loudspeaker in the 60cm arc.

The visual stimuli consisted of a virtual male face printed on cloth. This was done to make the faces acoustically transparent (McAnally & Martin, 2008; Perrott, Saberi, Brown, & Strybel, 1990). Each face was illuminated via computer controlled 5v LEDs. The auditory stimulus was a recording of a male voice saying “hi”. The duration for all stimuli was 200ms.

Design and Procedures

The experiment was based on a blocked 3x2x2x4 repeated measures design with Modality (vision only, auditory only, and audio-visual), Distance (60cm vs. 120cm), Spatial Congruency (congruent vs. incongruent), and Location (-67.6°, -22.5°, 22.5°, and 67.5°).

The Modality conditions were blocked and the order was counterbalanced across participants. In the vision only condition, a single face was illuminated. The participant’s task was to determine if the illuminated face was in the 60cm or 120cm arc. In the auditory condition, a single voice was presented and the participant's task was to determine if it emanated from the 60cm or 120cm arc. In the audio-visual condition, saw and heard a face. The participant's task was to ignore the auditory stimulus and indicate the distance of the visual stimulus. In 50% of trials the audio-visual stimuli were congruent. There were 80 trials in each condition.

Results and Discussion

A 2x2x4 repeated measures ANOVA was performed with Modality (vision vs. audition), Distance (near vs. far) and Location (-67.5, -22.5, 22.5, and 67.5) as factors. Participants were significantly more accurate when the audio-visual stimuli were spatially congruent compared to when they were incongruent [$F(1,24) = 4.95$, $p = 0.03$] (see Fig. 1). There was no significant difference between the AV Congruent and vision only conditions [$F(1,24) < 1$, *n.s.*] but between the AV Congruent and auditory only conditions [$F(1,24) = 37.04$, $p < 0.0001$].

Participants were also significantly faster when the audio-visual stimuli were spatially congruent compared to when they were incongruent [$F(1,21) = 4.08$, $p = 0.05$] (see Fig. 1).

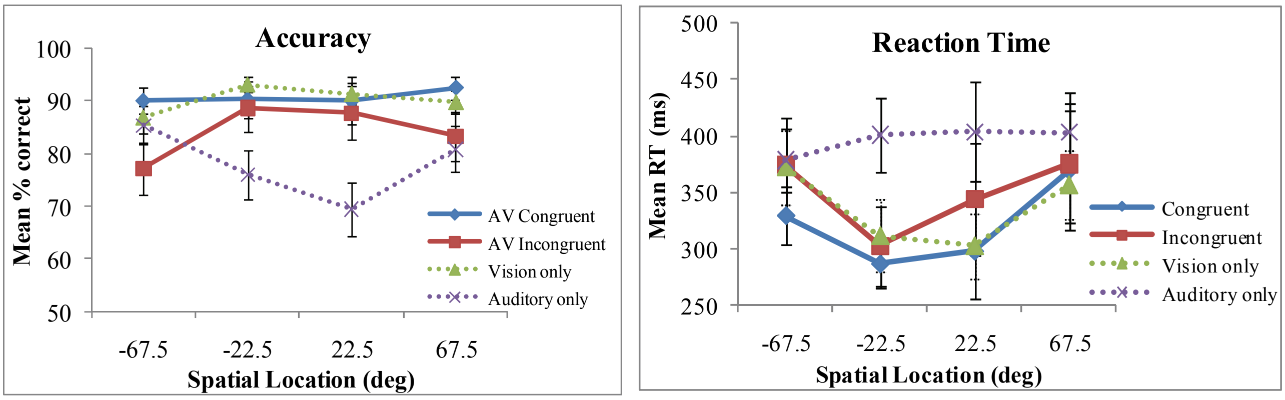


Figure 1. Accuracy and reaction time results for Experiment 1. Participants were significantly faster and more accurate and when audio-visual stimuli were spatially congruent in depth. Error bars represent the SEM.

Participants were more accurate to identify the depth of a visual stimulus compared to an auditory stimulus. Despite the relatively small spatial difference between the visual and auditory stimuli, participants were significantly more accurate and faster when audio-visual stimuli were presented in the same location compared to different locations. However, these differences were constrained to the peripheral locations. There was no significant effect for stimuli presented at ±22.5° of the central fixation. At ±67.5° from central fixation, accuracy was significantly reduced and reaction times were longer when audio-visual stimuli were presented at different locations.

There was a significant difference between the vision-only and AV incongruent condition, but not between the vision-only and AV congruent condition. This suggests that the difference between the AV congruent and incongruent was not due to multisensory facilitation, but due to multisensory distraction.

Experiment 2

The stimuli in Experiment 1 were relatively simplistic. There were very few locations and only one spatial difference between the auditory and visual stimuli. To explore the extent of which audition can provide a spatial context for which to locate a visual target we utilized virtual reality to increase the vary the distractor size and increase the number of spatial locations the objects can be placed.

Methods

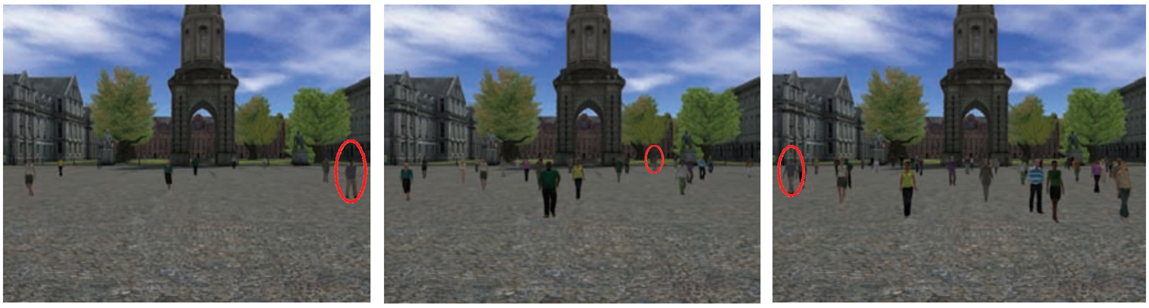
Participants

Thirty-five (19 = female) participants between the ages of 17 years and 40 years (mean age = 24 years) took part in this experiment. Participants had normal or corrected to normal vision and they did not report any hearing impairments.

Apparatus, Materials, and Procedures

Twenty-seven visual scenes of the Front Square of Trinity College Dublin were created in 3DStudio Max. Amongst the scene were randomly placed humanoid agents of which there was always one target agent (see Figure 2). There were three distracter sizes (8 agents, 26 agents, or 44 agents). The auditory stimulus was a male voice recorded in an open-air environment on a grass field. The recording distances were 1m, 2m, 5m, 10m, and 20m. These recordings were modulated in OpenAL to simulate the spatial depth. Generic HRTFs were used to accurately simulate external sound sources.

The visual stimuli were presented on a 22 inch widescreen computer monitor. The auditory stimuli were delivered through circum-aural headphones. The participants' task was to locate the visual target and indicate if it was placed in the left- or right-side of the scene. Participants were told to ignore the auditory stimulus.



9 Agents 27 Agents 45 Agents

Figure 2. Examples of three visual scenes with the difference distractor sizes. The target agent is circled. The target agent was always the same.

Results and Discussion

Participants were more accurate [$p < .05$] and faster [$p < 0.01$] when the auditory stimulus was spatially congruent to the visual stimulus (see Figure 3). Participants were also significantly slower [$p < 0.0001$] and less accurate [$p < 0.003$] as the number of agents increased.

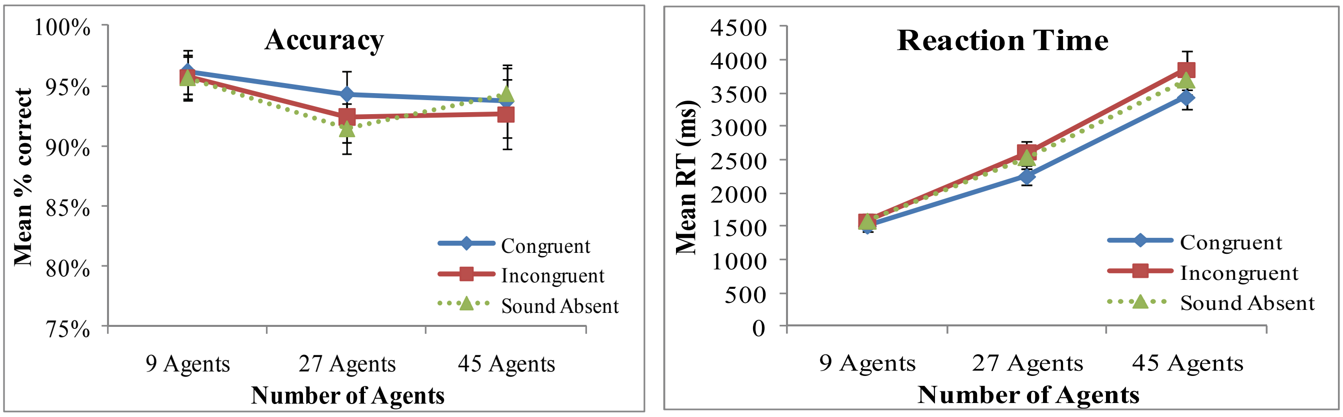


Figure 3. Mean accuracy and reaction times for the visual search task. Error bars represent the SEM.

General Discussion

The importance of congruent depth information between vision and auditory stimuli are clearly illustrated by these results. Spatially incongruent auditory stimuli reduced accuracy and increased reaction times for identifying the location of the visual target. This extends the findings of previous studies (Perrott, Cisneros, McKinley, & D'-Angelo, 1996; Perrott, et al., 1995) who show an improvement in accuracy and reaction times when the auditory and visual target are presented in the same location across the horizontal plane. It is important to note that those previous studies found a multisensory facilitation that was significantly better than detecting the target using visual alone.

The audiovisual effects in Experiment 1 were isolated to the peripheral locations. There was no multisensory effect in the central locations. The auditory stimuli reduced localization performance in the peripheral locations where spatial identification of the visual targets was more difficult. This suggests that participants will ignore the auditory stimulus when there is a reliable visual signal. However, in the more ‘noisy’ peripheral visual signals auditory information is utilized.

The results of Experiment 2 demonstrate that virtual audio can provide a spatial cue to find visual targets in a 2-dimensional virtual environment when contextually relevant. Participants used the auditory cue as an indicator to where to search in the scene. It is interesting to note that the entire scene was within a smaller visual angle than between the two central loudspeakers of Experiment 1. The audiovisual enhancement in Experiment 2 suggests that the task in Experiment 1 did not require participants to use the auditory cue in the area directly in front of them. It is possible that performance would improve at those locations for the congruent condition compared to the incongruent condition if the task was more difficult.

In conclusion, we demonstrate that it is necessary to have audiovisual stimuli congruently placed in depth as well as in the horizontal plane. This effect is can be generalized to a virtual reality. Audio cues guide the participant's eyes to the most likely spatial location in the scene.

Acknowledgements

This research was funded by Science Foundation Ireland as part of the METROPOLIS project.

References

Gardner, M. B. (1968a). Proximity image effect in sound localization. *Journal of the Acoustical Society of America, 43*, 163.

Hirsh, I. J., & Sherrick, C. E., Jr. (1961). Perceived order in different sense modalities. *Journal of Experimental Psychology, 62*(5), 423-432.

Lewald, J., & Ehrenstein, W. H. (1998). Auditory-visual spatial integration: A new psychophysical approach using laser pointing to acoustic targets. *Journal of the Acoustical Society of America, 104*(3), 1586-1597.

Lewald, J., & Guski, R. (2004). Auditory-visual temporal integration as a function of distance: No compensation for sound-transmission time in human perception. *Neuroscience Letters, 357*, 119-122.

McAnally, K., & Martin, R. (2008). Sound localisation during illusory self-rotation. *Experimental Brain Research, 185*(2), 337-340.

Perrott, D. R. (1984). Discrimination of the spatial distribution of concurrently active sound sources: Some experiments with stereophonic arrays. *Journal of the Acoustical Society of America, 76*(6), 1704-1712.

Perrott, D. R., Cisneros, J., McKinley, R. L., & D'-Angelo, W. R. (1996). Aurally aided visual search under virtual and free-field listening conditions. *Human Factors, 38*(4), 702-715.

Perrott, D. R., Cisneros, J., McKinley, R. L., & D'Angelo, W. R. (1995). Aurally Aided Detection and Identification of Visual Targets. *Human Factors and Ergonomics Society Annual Meeting Proceedings, 39*, 104-108.

Perrott, D. R., Saberi, K., Brown, K., & Strybel, T. Z. (1990). Auditory psychomotor coordination and visual search performance. *Perception & Psychophysics, 48*(3), 214-226.

Sugita, Y., & Suzuki, Y. (2003). Audiovisual perception: Implicit estimation of sound-arrival time. *Nature, 421*(6926), 911-911.

Zampini, M., Shore, D. I., & Spence, C. (2003). Audiovisual temporal order judgments. *Experimental Brain Research, 152*, 198-210.