AURALLY AIDED VISUAL SEARCH WITH MULTIPLE AUDIO CUES

Brian Simpson and Nandini Iyer

Air Force Research Laboratory
Wright-Patterson Air Force Base, OH
brian.simpson@wpafb.af.mil
nandini.iyer@wpafb.af.mil

Douglas S. Brungart

Army Audiology and Speech Center
Walter Reed Army Medical Center
Washington, DC
douglas.brungart@us.army.mil

ABSTRACT

In many applications, the primary goal of a spatialized audio cue is to direct the user's attention to the location of a visual object of interest in the environment. This type of auditory cueing is known to be very effective in environments that contain only a single visual target. However, little is known about the effectiveness of this technique in environments with more than one possible target location. In this experiment, participants were asked to identify the characteristics of a visual target presented in a field of visual distracters. In some conditions, a single auditory cue was provided. In other conditions, the auditory cue was accompanied by one or more audio distracters at different spatial locations. These conditions were compared to a control condition in which no audio cue was provided. The results show that listeners can extract spatial information from up to three simultaneous sound sources, but that their visual search performance is significantly degraded when more than four simultaneous sounds are present in the stimulus.

1. INTRODUCTION

In many practical applications of virtual audio displays, the primary purpose of the spatialized auditory cue is to direct the user's attention to the location of a target object so a positive visual identification can be made. In situations where the visual field is cluttered and the target object is difficult to distinguish from other visual objects in the environment, dramatic reductions in visual search time can be achieved simply by turning on a broadband sound at the location of the target. For example, Bolia et al. [1] examined the benefit of audio cueing as a function of visual scene complexity by manipulating the number of visual objects in the scene (i.e., the set size). In a two-alternative, forced-choice task, the subjects were to detect and identify which of two target light arrays was presented on each trial. They found that, when no audio cue was presented, visual search times increased with increasing set size, consistent with a limited-capacity attentional process in which an observer must serially scrutinize each display element individually. However, when an audio cue was presented from the location of the target, response times were significantly reduced relative to the no-cue condition (by up to 93%), and were essentially independent of set size, suggesting that the benefit of providing an auditory cue that is spatially coincident with a visual target not only reduces target acquisition times dramatically, but in fact changes the nature of the search strategy. Specifically, the salience of the auditory cue leads to searches that are more characteristic of parallel search processes, and thus are essentially independent of set size.

From these results, it is evident that a continuous spatialized audio cue at the location of the target is almost always an advantageous display strategy in cases where the listener's visual attention should be directed to a single known location in space. However, the situation gets more complicated in cases where it is necessary to cue more than one target location at the same time. This might occur because the range of possible target locations has been narrowed down to one of N possible locations, or it might occur because more than one simultaneous target exists and the relative priority of each target cannot be determined without visual inspection by the operator. In either case, care must be taken in determining how to provide spatial auditory cues at the location of more than one simultaneous target. The simplest strategy is to turn on a different independent continuous sound source at each potential target location. However, each additional simultaneous sound source will reduce the localizabilty of the individual sources in the mixture [2], and as a result one would expect the advantage of audio cueing to decrease as the number of cued target locations increases. In the limit, one would expect performance to deteriorate to the point where no measurable advantage in search time is observed from the addition of spatialized audio cues at the locations of the potential targets.

In this experiment, the aurally-aided visual search paradigm employed by Bolia et al. [1] has been adapted to examine how visual search times change as a function of the number of auditorally-cued *potential* target locations within a set of 50 visual distracters. The next section describes the experimental procedures in more detail.

2. METHODS

2.1. Apparatus

The experiments were conducted in the Auditory Localization Facility (ALF) at Wright-Patterson AFB in Dayton, Ohio (Figure 1). The ALF is a geodesic sphere 4.3 m in diameter that is equipped with 277 full-range loudspeakers spaced roughly every 15° along its inside surface. The ALF facility is connected to a high-powered signal switching system that allows up to 16 different sounds to be routed to any or all loudspeakers from a multichannel digital soundcard (RME).

Mounted in front of each loudspeaker in the ALF facility is a small visual display consisting of a cluster of four red LEDs arranged in a square pattern, with each diode subtending a visual angle of approximately 0.5°. Figure 2 shows an illustration of the possible modes of these LEDs.

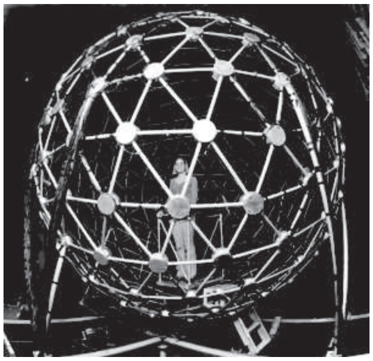

Figure 1: Auditory Localization Facility used for HRTF collection

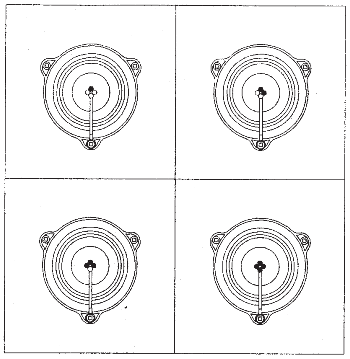

Figure 2: Configurations of LEDs used for visual display in experiment. LED clusters with an odd number of active LEDs (1 or 3) were used as visual distracters (left column). An LED cluster with an even number of active LEDs (2 or 4) was used to designate the target location. Participants were required to search all the loudspeaker locations to find the location with an even number of active LEDs, and then press a button to identify whether the target had 2 or 4 LEDs active.

2.2. Participants

A total of 8 paid volunteer listeners participated in the experiment, including 4 males and 4 females. All had normal audiometric thesholds, and their ages ranged from 19 to 25 (Mean age 23 years). All were screened to have uncorrected 20/20 vision in both eyes.

2.3. Procedure

The experiment was conducted with participants standing on a platform in the center of the ALF facility. The participants wore a headband with a 6-DOF headtracking sensor (IS-900) attached, and, at the start of each trial, they were asked to turn and face the front speaker in the ALF facility until an LED cursor slaved to the participants' head orientation was activated at that loudspeaker. They then pressed a button to indicate their readiness to begin the trial. At that point, two things happened. First, a visual display was generated by randomly selecting one loudspeaker as the target loudspeaker and turning on either two or four LEDs at that loudspeaker location, and then randomly selecting 8, 16, or 50 other loudspeaker locations as ”visual distracters” and turning on 1 or 3 LEDs at each of those loudspeaker locations (see Figure 2). Second, a broadband continuous noise signal was switched on at the location of the target, and additional, statistically-independent random noise signals were simultaneously switched on at 0, 1, 2, 3, 5, 7, or 15 other audio distracter locations. These audio distracter locations were chosen randomly from among the locations of the visual distracters. Thus, in the condition with 50 visual distracters and 8 audio distracters, the 277 speakers in the ALF facility included: one *target* speaker with a continuous noise signal and either 2 or 4 active LEDs; seven *audio distracter* loudspeakers with a continuous noise signal and either 1 or 3 active LEDs; and 43 *visual distracter* loudspeakers with no sound but either 1 or 3 active LEDs.

In all cases, the participant's task was the same: search all the loudspeaker locations with active sound sources for the target location with an even number of active LEDs (2 or 4), and press a response button to indicate whether there were 2 or 4 LEDs active at the target location.

Responses were collected in blocks of 20 trials. On each trial, the number of audio distracters and visual distracters was randomly chosen. Most of the data were collected in conditions with 50 visual distracters. Over the course of the experiment, each of the eight participants provided responses in 60 trials in conditions with 50 visual distracters and 0, 1, 2, 3, 4, 5, 7 or 15 auditory distracters. They also participated in three visual-only control conditions with 8, 16, or 50 visual distracters but no auditory signals. In total, a minimum of 700 trials were collected on each of the eight participants in the experiment.

3. RESULTS

Listeners were instructed to conduct the task as quickly as possible while ensuring a very high level of accuracy on the identification of the number of LEDs at the target location. As a result, overall accuracy on the LED identification task was extremely high- listeners correctly distinguished between target configurations containing 2 or 4 LEDs in 99.82% of all trials.

The more meaningful metric of performance in the task is response time, measured from the presentation of the stimulus at the beginning of the trial to the time when the participant pressed the response button identifying the target, which terminated the trial. Figure 3 shows performance as a function of the number of visual distracters in the visual-only control condition, where no audio stimulus was presented, averaged across all participants. As would be expected, the amount of time required to complete the task increased systematically as the number of visual distracters increased, suggesting a serial search process. When 50 visual distracters were present, response time was on average about 8 seconds.

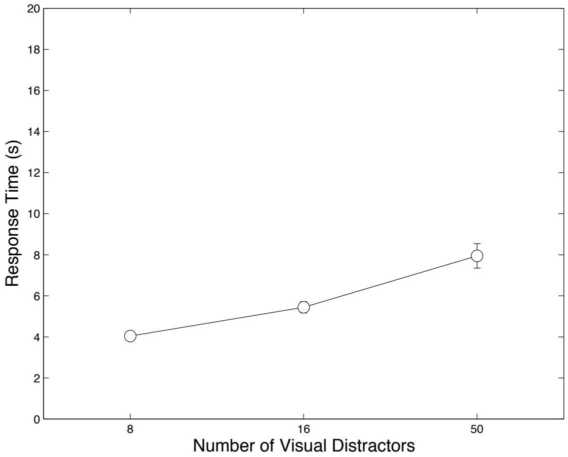

Figure 3: Response times, averaged across all participants, plotted as a function of the number of visual distracters in the visual-only control condition. The error bars show the 95% confidence intervals around each data point.

Figure 4 shows mean response time in the experiment as a function of the number of audio distracters in the conditions with 50 visual distracters. For comparison purposes, the shaded bar in the middle of the figure shows performance in the visual-only condition with 50 visual distracters and no audio stimuli. Again, as expected, the overall visual search time was found to increase with the number of audio distracters. Moreover, as expected, the benefit of having an audio signal at the location of the target disappears after the addition of a certain number of audio distracters. Specifically, there is no longer a difference between the visual-only condition and the audio condition when three audio distracters are added to the stimulus.

What is somewhat surprising about the data, however, is that performance does not merely plateau when enough audio distracters are added to the stimulus to eliminate any useful information the listener might obtain from the audio cue at the location of the target. Rather, it continues to worsen, and when 15 audio distracters were present, the total search time to find the target was almost twice as long as it was when no audio signals were presented at all. Importantly, this result suggests that listeners are not generally able to determine when audio information no longer provides any advantage in this visual search task. In such cases, one might expect that the participants would adjust their strategy and ignore this distracting audio information. Rather, the fact that response times continue to increase with the number of sounds suggests that participants are searching serially through the sounds in order to locate that sound associated with the visual target. This means that audio display designers must use extreme caution when they implement audio displays that have the potential to generate large numbers of spatialized cues at the same time. The results of this experiment suggest that the users of these systems may not be able to accurately determine when the audio information should be relied upon for a visual search task, and when it should be ignored. Consequently, it seems that there may be some cases where the provision of additional audio information might actually significantly degrade the operator's performance in complex visual search tasks.

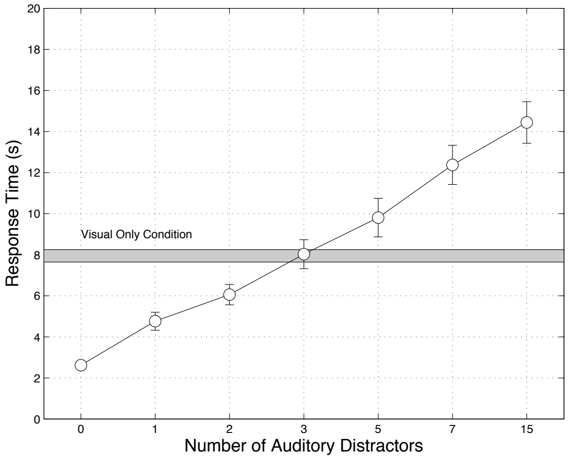

Figure 4: Response times, averaged across all participants, plotted as a function of the number of audio distracters in the conditions with 50 visual distracters. The error bars show the 95% confidence intervals around each data point. The shaded bar shows mean performance in the visual-only condition with 50 visual distracters.

4. CONCLUSIONS

In this experiment, we examined how well participants were able to perform a complex aurally-aided visual search task when one or more distracting sounds were presented concurrently with the audio cue from the location of the visual target. The experiment was intended to replicate the kind of scenario that might occur when an operator is required to investigate more than one simultaneous visual target, or when a visual target or threat is known to be present at one of a small number of possible locations. In cases where there are fewer than four simultaneous targets, these results suggest that some advantage can be gained simply by providing a co-located continuous sound source at all the possible locations in the target set. However, when more than four target locations need to be cued, the presentation of simultaneous co-located audio cues at the target locations actually results in a significant *degradation* in performance relative to the visual-only case where no cueing sounds are provided.

However, it is important to note that these results only apply to the worst-case condition where the exact same audio cue is provided at all the possible locations in the target set. While there is no guarantee that performance would be improved by other types of cueing sounds, it is likely that some alternative audio symbology incorporating sounds that do not overlap either in time or frequency might be able to produce better performance in this task than was obtained with the continuous broadband noises used in this study. We are currently conducting experiments to explore this possibility in more detail.

5. REFERENCES

[1] R. Bolia, W. D'Angelo, and R. McKinley, "Aurally aided visual search in three dimensional space," *Human Factors*, vol. 41, pp. 662–669, 1999.

[2] D. Brungart and B. Simpson, "Within-ear and across-ear interference in a dichotic cocktail party listening task: Effects of masker uncertainty," *Journal of the Acoustical Society of America*, vol. 115, pp. 301–310, 2004.