AN AUDITORY TWIST ON VISUAL SEARCH USING HEAD MOVEMENT CONTROLLED AMBISONICS

Kat Agres

Cornell University
Department of Psychology

Spencer Topel

Dartmouth College
Digital Musics

Michael Spivey

University of California, Merced
School of Social Sciences,
Humanities, and Arts

Stephen Moseson

Cornell University
Sibley School of Mechanical
Engineering

ABSTRACT

In this paper, we propose an auditory search task using a virtual ambisonic environment presented through static Head-Related Transfer Functions (HRTF's). Head-tracking using a magnetometer captures the listener's orientation and presents an interactive auditory scene. Reaction times from 15 participants are compared for Simple and Complex auditory search tasks. The results lend support to the hypothesis that similar attentional mechanisms may constrain processing during visual and auditory search tasks.

1. INTRODUCTION

Visual search tasks have elucidated fundamental properties of visual perception, such as attention and efficiency of processing [1, 2]. When performing a simple visual feature search, such as finding a blue line amongst red lines, the target is found rapidly regardless of the number of distractors. This is called the "pop-out effect".

A more challenging search task is the conjunction search, in which two or more features are required to identify a target. Finding a target consisting of a conjunction of features often requires greater attentional resources and more time, and search time increases with the number of distractors present. In visual search tasks, features such as color, orientation, and shape are often manipulated.

The present study is a novel comparison to auditory perception using auditory analogs of visual search features, namely, pitch and timbre. Due to our novel methodology, we are also able to test spatial distribution, as performed in both auditory and visual search tasks [11].

Designing an experiment to include spatial distribution presents a unique challenge because listeners cannot actively use visual resources in the identification process, i.e. through a visual interface. Based on [12], we employed an active third-order ambisonics scene using ICST Zurich's Ambisonics Toolbox through what the authors coined as "time-invariant" HRTFs. Presentation in this manner allowed for spatial movement to share a direct relationship with head movement and to limit the interactive listening environment to a reproducible space [3].

There has been a debate in the psychological literature on attention as to whether attentional mechanisms are shared across modalities. Therefore, this experiment sought to use an auditory task analogous to a classic visual task that measures efficiency of processing and attention. We hypothesized that attention is similarly limited in both the auditory and visual modalities when more than one feature must be attended in a search task. Therefore, we predicted that listeners would require more time for complex auditory searches than for simple searches, akin to findings in visual perception. To this end, we measured reaction time, the amount of time that each listener took to identify a target sound among distractor sounds. Our results support connections between auditory and visual processes illustrated in prior research, especially Albert Bregman's research on auditory scene analysis [4].

2. BACKGROUND AND PURPOSE

2.1. Auditory Feature Integration

Feature Integration Theory pioneered by Anne Treisman [2], frames perception of a visual scene through a dialectical relationship between targets and distractors; much the same way information and noise operate over transmission channels in Information Theory [5].

In particular interest to this study is the conjunction search, where two or more features are required to identify a target. In these types of tests, reaction time increases directly with the number of distractors present. Therefore, conjoining multiple features requires additional processing, which in turn produces an inefficient search (see Figure 1).

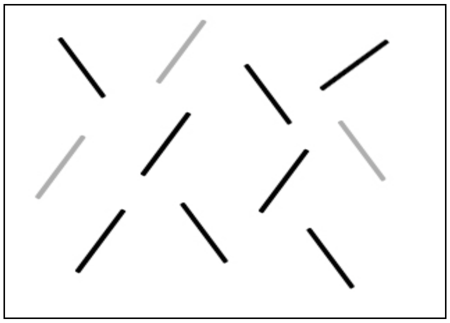


Figure 1. In this visual perception task, the slight shift in angle of the character in the top-right corner is the target, while the other shapes are distractors.

Although some research has investigated the perception of spatially distributed auditory objects [6], few studies have utilized a search task paradigm to test whether phenomena such as the "pop-out effect" on tasks requiring spatial attention resources are exclusive to visual perception or are present across modalities.

Testing attentional constraints in the auditory domain allows us to determine whether the properties of spatial attention are solely determined by a single modality, or result from a more domain general process.

2.2. Technical Rationale

The virtual Ambisonic approach is built on the rationale that headphones, through HRTF's, can present an accurate auditory scene [12]. We propose a head-tracking system that combines software interaction in the Max 5 environment with a USB capable solid-state 3 axis digital compass [9], fixed directly on top of a pair of headphones that allows us to present auditory stimuli to a listener with minimal user interface.

The goal of the head-tracking system was to create an interactive sound field where sounds move depending on the listeners' head position. Our spatial coding was restricted to a single listening plane, and therefore movement was invariant to vertical head motion and tilt information.

To build upon hands-free "selection", we developed the idea of a sound point being "in focus" when the user looks in the direction of a sound source. A boundary angle, represented by the shaded area in Figure 2, activates the focus effect drawing the sound source "toward" the listener and filling the listening space. When not in focus, sound points move back into an amplitude-congruent listening space.

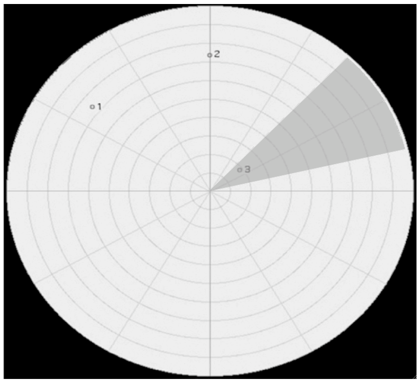


Figure 2. The ICST Ambimonitor object used in the experiment with three points indicating virtual sound sources. The center of the graph indicates the listeners' immediate headspace, and the proximity of the third point shows that point as being "in focus". The shaded area shows the listener's head direction and focus area.

2.3. Implementation

To move points in the experimental listening space, the Max/MSP object Ambimonitor, Figure 2, was employed with our position calculations. Since the Ambimonitor uses a navigational coordinate system, we implemented our position calculations using linear distance functions to account for point distribution, focused, and unfocused movement.

Focused distance was calculated using the inverse exponential function and unfocused distance was determined by a monotonic increasing function. The positional control units in the program are measured in dB, and then converted into relative distance. In our experiment, we used a distance factor of 6, where one unit of relative distance corresponds to a 0.3 dB difference in amplitude [6].

The actual position of each point is smoothed using a first order differential equation:

$$\tau \bullet \dot{x}(t) + x(t) = y(t) \tag{1}$$

where $x(t)$ is the actual point-distance given to the ambiencode object, $\dot{x}(t)$ is the derivative of the actual point distance with respect to time, $y(t)$ is the input point distance , and τ is the time constant. In the program τ = Ramp Time in milliseconds.

3. METHOD

3.1. Experimental Environment and Design

Fifteen undergraduates participated in the study for extra credit in a psychology course, and all had normal hearing. The experimental session took place in a well-lit laboratory room, and the participants were run one at a time. Each participant was seated in front of a MacBook Pro laptop computer, the task was explained, and then the participant put on the magnetometer-mounted headphones. The computer screen served to give participants a starting fixation point (looking straight ahead), and the magnetometer was manually calibrated for each participant at the beginning of the experiment.

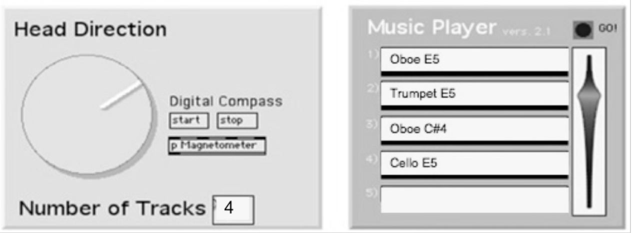
All of the tones (sound sources) in the study were created using Finale music software. These tones were looped continuously throughout every trial. Each tone was sampled from a different Finale MIDI instrument, and featured distinct spectral characteristics. Although presenting continuous tones diminished their attack and decay, the spectral content of each tone will henceforth be referred to simply as “timbre” for simplicity.

Before starting the experimental trials, participants first completed a three-trial practice session to familiarize themselves with the technology and task requirements. The experimental session consisted of 84 trials. At the beginning of every trial, the target tone was played, which consisted of a particular timbre (cello or flute) and pitch (C#4 or E5). After hearing the target, the listener was presented either two, three, or four tones in distinct auditory locations within the front hemisphere of space. In half the trials, the target was not present. In the other trials, the location of the target was randomized, along with one, two, or three distractor (non-target) tones. Each distractor was either a C#4 or E5, and featured an oboe, trumpet, cello, or flute timbre.

The listeners’ task was to locate the target among the distractors in the sound space in front of them. The distractors could share either the same pitch or timbre as the target, or consist of different features. To “find” the target in space, the participant moved his or her head left and right, and the direction of gaze was recorded by the magnetometer.

Looking around the sound space altered the position of the constituent tones of the trial such that the tones located in the space surrounding the direction of gaze (i.e., within the focus angle) increased in presence, while the other tones perceptually receded into the distance. Participants looked around this interactive auditory scene until they found the target tone; they recorded their response (the location of the target) by gazing towards the target and pressing a button on the keyboard. For trials in which no target was present, listeners could press the space bar to log a “no target present” response.

Analogous to some visual search tasks, the search type on each trial could either be Simple or Complex. In Simple searches, the distractor(s) did not have either feature (timbre or pitch) in common with the target. In Complex searches, one of the distractors exhibited either the same timbre *or* pitch as the target. If, for example, the target was a cello E5 tone, then one of the non-target distractors was either a cello C#4 tone, or an oboe, trumpet, or flute tone with the pitch E5. Reaction time was recorded on each trial to quantify the time required to find the target among distractors. Accuracy of response was also recorded.



Number of Tracks 4

1) Oboe E5
2) Trumpet E5
3) Oboe C#4
4) Cello E5
5)

Figure 3. Experiment interface for experimentor, calibrated before the start of the experiment.

4. RESULTS AND DISCUSSION

4.1. Results

The results provide evidence that listeners do require more time for complex searches than for simple searches. A 2 X 3 ANOVA of Search Type (Simple or Complex) X Set Size (the number of simultaneous tones per trial) yielded a significant main effect of Search Type (F = 5.84, p < .05), with Complex search eliciting longer reaction times than Simple search. There was also a significant main effect of Set Size (F = 6.75, p < .01), with reaction times increasing with the number of distractors. The interaction of Search Type and Set Size was not significant (F = 1.13, p = .3).

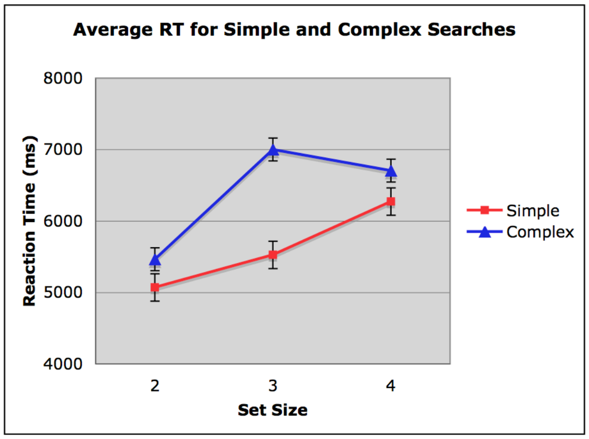


Figure 4. Average reaction time for Simple and Complex searches where Set Size is the number of simultaneous tones per trial.

Separate analyses were then run to isolate main effects within each search type. Simple search did not exhibit a significant effect of Set Size on reaction time (F = 2.41, p < .1), whereas complex search showed a robust and highly significant effect of Set Size (F = 5.84, p < .01), as shown in the Figure (4). A contrast confirmed that the Set Size of 2 tones yielded significantly lower reaction times than a set of 3 or 4 tones (F = 11.61, p < .01). Accuracy was lower for Complex search (73%) than for Simple search (83%)

and was reduced slightly as Set Size increased, thus providing no indication of a speed-accuracy tradeoff.

4.2. Discussion of Results and Methodological Concerns

The behavioral findings support the hypothesis that similar attentional constraints exist for auditory and visual searches. In Complex searches, the listener is required to focus on more than one acoustic feature because one of the distractors shares a feature in common with the target. This eliminates the likelihood of an auditory "pop-out" effect, and the search time is inflated, especially when more than one distractor is present. This result is consistent with the literature on visual search [2]. For Simple searches, reaction time increased somewhat with Set Size, but this effect was not statistically significant. This lack of a significant effect of set size on Simple auditory search comports with findings in visual search.

In the future, a greater number of distractors should be utilized to explore whether this trend becomes significant (the slope continues to increase significantly), or the reaction time levels off. Another possible reason for the increase in reaction time for the Simple search task is that a larger set size may inherently require a longer reaction time, simply because more time may be needed to look around the spatial field for each tone.

This experiment diverges from traditional visual search tasks because there is almost always more than one distractor that shares a feature in common with the target in Complex visual searches. Usually at least one distractor will be present for each stimulus feature; in other words, the set of distractors in visual search tasks contain all of the target's features, (i.e. if the target is a blue "L", one distractor would be a blue "I" and another would be a green "L"). In our experiment, only one feature was shared between the target tone and the other tones (either Pitch *or* Timbre on each trial). Inclusion of distractor sets containing both features may result in a larger difference in reaction times between the Simple and Complex searches.

Another consideration is that additional attentional resources may be required to process and remember the spatial position of the tones. This may account in part for the greater reaction times found in our study as compared to typical reaction times in visual search tasks, which are an order of magnitude smaller.

4.3. Future Research

New experiments with auditory spatialization might include adjustments to the spatial field including size and angle projections to account for discrepancies in reaction times. It would be interesting to consider the hypothesis that additional attentional resources may be required to process and remember the spatial position of distinctive auditory sources, as suggested in a study that considers "multi-sensory convergence of spatial awareness" [10]. Also, we plan to test the effect of the order of ambisonics as well as plain HRTF's on the subjects' performance in this task.

Additionally, future work should include more than one distractor in the Complex search condition to allow for a more analogous comparison of auditory and visual processing.

Materials to run this experiment are available at http://digital.music.cornell.edu

5. REFERENCES

Acknowledgement of Research Funding
Supported by NSF grant BCS-0721297 to MJS

[1] Duncan, J. and G. W. Humphreys (1989). "Visual search and stimulus similarity." Psychological Review **96**(3): 433-458.

[2] Treisman, A. (1982). "Illusory conjunctions in the perception of objects." Cognitive Psychology **14**(1): 194-214.

[3] Gardner, W. G. and K. D. Martin. (1995). "HRTF measurements of a KEMAR." J. Acoust. Soc. Am. **97(6)**: 3907-3908.

[4] Bregman, A. S. (1990). Auditory scene analysis: the perceptual organization of sound, MIT Press.

[5] Healey, C. G. (May 11, 2009). "Perception in Visualization." From http://www.csc.ncsu.edu/faculty/healey/PP/index.html

[6] Pakarinen, S., R. Takegata, et al. (2007). "Measurement of extensive auditory discrimination profiles using the mismatch negativity (MMN) of the auditory event-related potential (ERP)." Clinical Neurophysiology **118**(1): 177-185.

[7] Schacher, J. C. and P. Kocher (2006). "Ambisonics spatialization Tools for Max/MSP." International Computer Music Conference: 274-276.

[8] Gardner, W. G. and K. D. Martin. (1995). "HRTF measurements of a KEMAR." J. Acoust. Soc. Am. **97(6)**: 3907-3908.

[9] OceanServer Technology, I. (2003-2010). Retrieved January 7, 2010, from http://www.ocean-server.com/compass.html.

[10] Clavangnier, S. (2004). "Long-distance feedback projections to area V1: Implications for multisensory integration, spatial awareness, and visual consciousness." Cognitive, affective, & behavioral neuroscience 4(2): 117.

[11] Eramudugolla R., K. McAnally, et al. (2008). "The role of spatial location in auditory search" Hearing Research **238**: 139-146

[12] Noisternig M., T. Musil, et al. (2003). "3D Binaural Sound Reproduction using a Virtual Ambisonic Approach" VECIMS Lugano, Switzerland.