Searching for Sources from a Fixed Point in a Virtual Auditory Environment

Agnieszka Roginska[1], Gregory H. Wakefield[2], Kyla McMullen[2]

[1]Music and Audio Research Lab, New York University, 35 West 4th St, New York, NY 10012

[2]Computer Science and Engineering Division, Department of Electrical Engineering and Computer Science, The University of Michigan, Ann Arbor, MI 48109

roginska@nyu.edu, ghw@umich.edu, kyla@umich.edu

ABSTRACT

Interaction between the listener and their environment in a spatial auditory display plays an important role in creating better situational awareness, resolving front/back and up/down confusions, and improving localization. Prior studies with 6DOF interaction suggest that using either a head tracker or a mouse-driven interface yields similar performance during a navigation and search task in a virtual auditory environment. In this paper, we present a study that compares listener performance in a virtual auditory environment under a static mode condition, and two dynamic conditions (head tracker and mouse) using orientation-only interaction. Results reveal tradeoffs among the conditions and interfaces. While the fastest response time was observed in the static mode, both dynamic conditions resulted in significantly reduced front/back confusions and improved localization accuracy. Training effects and search strategies are discussed.

1. INTRODUCTION

Static positioning of the listener within a virtual auditory environment (VAE) presented over headphones can limit the degree to which the listener correctly identifies the spatial relationships among acoustic sources in that environment. Allowing the listener to move through the environment appears to resolve such ambiguities [1]. When asked to walk to the position of a source within a VAE, listeners are able to do so with the aid of a head-tracking system. Learning to navigate the VAE with a head-tracker requires a minimal amount of training. Once trained for a particular VAE and tracking device, listeners appear to maintain their ability to navigate rapidly. Similar behavior is observed when listeners navigate the VAE using a mouse-keyboard-display interface. Listeners are able to quickly learn to move the position of an avatar on a visual display to the location of a source and to do so with the same level of accuracy as observed for walking to the source location.

In both the cases of walking through an environment and moving an avatar through a visual display of the environment, perceptual ambiguities are resolved by changes in the acoustic field based on the position of the listener relative to the positions of the sources. These include changes in the binaural cues based on the orientation of the head relative to the sources as well as changes in the monaural cues based on the relative distances between the listener and the sources. The present study asks whether listeners can accomplish the same task through changes in orientation alone. From a practical standpoint, it is not always feasible to have listeners "walk" through a VAE in order to generate a sufficient amount of dynamically-varying acoustic information to learn where sources are located. A simple turn of the head is known to provide sufficient cues to resolve front-back confusions in localization of sources in an anechoic environment [5][6]. Such is expected to be sufficient for localizing sources in azimuth in a VAE by monitoring head position using a head tracker. While we may expect that a similar rotation of an avatar's head is sufficient, the extent to which accuracy and search times are affected by mediating listener orientation using an avatar rather than head motion is important in the design of human-computer interfaces for VAE.

2. METHODS

Based on the methods and procedures of studies performed in the earlier stages of this line of experiments [1][2], an experiment was designed to measure human performance during an auditory search task. We studied the effect of the type of interface used by the participants to interact with the auditory environment. Three conditions of interaction between the participant and the environment were studied: static (no interaction), mouse interaction (avatar mediation), and tracker interaction (natural mediation). A single acoustic source was positioned in a virtual anechoic auditory environment along the horizontal plane. Participants were asked to locate the source, and mark its location on a visual display.

During the static condition, participants did not interact with the environment. The source was presented at an absolute, fixed location in relationship to the listener. Listeners had to judge the absolute location of the source.

In the avatar mediation, participants used the mouse interface to indirectly change their relative position to the environment, and sources within this environment, by orienting the "nose" of the avatar. It was possible to change only the orientation of the avatar, and not its position. As the orientation of the avatar was updated, the relative position of the source to the avatar was calculated and the audio was processed to reflect this relative change in position.

A head tracker was used to support natural mediation of head orientation. The orientation information (yaw, pitch and roll) was captured and used to calculate the relative position of the source to the listener. The relative position of the source was calculated such that the location of the source appeared to

be constant within the environment, as the listener turned their head.

During the three-stage listening experiment, participants first went through a training phase for each interaction condition, followed by the testing phase.

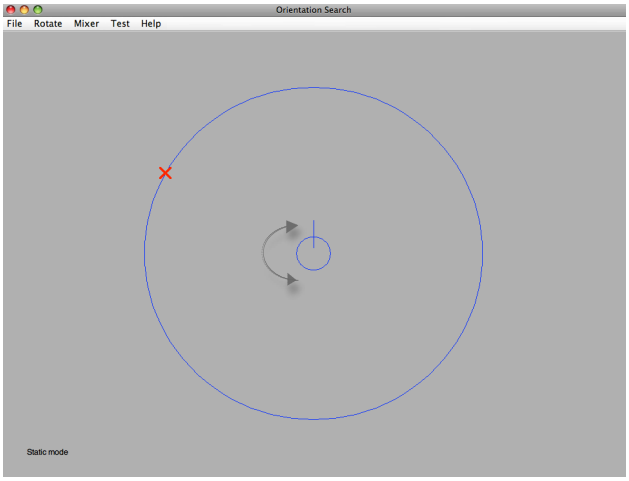

Figure 1: GUI used by subjects to mark location of target source. The target location is marked with an 'X' along the horizontal plane circle. The orientation of the listener's perspective is represented by the direction of the "nose" in the center of the figure.

2.1. Procedure

The experiment began by informing subjects that the task they must complete involved locating a single source along the horizontal plane, and marking the location of this source on the GUI, such as the one in Figure 1.

At the beginning of the session, the subject was randomly assigned the order in which the mediations would be presented. Each mediation consisted of a training phase followed by a testing phase. For example, a subject might do the experiment in the following order: training-static, testing-static; training-avatar, testing-avatar; training-natural, testing-natural. The purpose of training phase was to familiarize subjects with the interface and the task up to a baseline measure of learning. Subjects ran a minimum of 10 training trials. If the standard deviation of the search times for the last four trials was less than 2.5 seconds, it was said that the subject achieved asymptotic, or optimal, performance, training was terminated, and the testing phase begun. If, however, the standard deviation was larger than 2.5 seconds after 10 training trials, additional runs of four trials were presented, until the search time of the latest four consecutive trials had a standard deviation of less than 2.5 seconds. Note that the measure of familiarization is based solely on search times: search accuracy was not used to determine whether a subject was sufficiently trained in the use of the interface.

At the onset of every trial, in the training as well as the testing phases, participants were first presented with a 4-second cue – a sample of the source they would be attending to. This cue was presented diotically. Following the cue, the target source was positioned at a random azimuth on the 0˚ elevation plane, and the participants began their search. The source was presented at a fixed distance from the listener.

When a participant finished both the test and training phases for one modality, they repeated the procedures for the other two modalities.

2.2. Apparatus

The test was conducted in the Spatial Audio Research Lab in the Music Technology program at New York University. Subjects were seated in front of a 17" monitor displaying the GUI shown in Figure 1. Sennheiser HD650 headphones were used to present stimuli.

In the static mediation, participants did not interact with the environment, and the location of the target source was fixed. The avatar mediation used the mouse as the interface to change the orientation of the perspective of the listener. The location of the mouse on the GUI determined the orientation of the "nose" of the avatar. The orientation of the listener's perspective was reflected by a graphical representation of the avatar's orientation. The natural mediation was controlled by the 6DOF Polhemus Liberty tracker. The sensor was mounted on the top of the headphones worn by the participant. Only the yaw information was collected and used to process the relative location of the source to the listener, all other position and orientation information was ignored. The yaw information was sampled at 10Hz and used to drive the signal processing engine.

A real-time spatial audio processing engine was developed in Matlab by using the Psychtoolbox extension of OpenAL. Head-Related Impulse Responses (HRIRs) measured on KEMAR at the NSMRL facility were used to process the spatial sound [3]. The signal processing was implemented in the same manner as in prior test in this line of experiments [1][2].

2.3. Sources

Four sounds were selected from the publically available BBC Sound Effects Library [4]. Sounds included a typewriter, a brook, crowd noise, and electronic music. During each trial, one of the sources was randomly selected. The sounds were 24-60 seconds in duration with continuous signal. Stimulus levels were adjusted to achieve equal sensation level. The adjustment was made by one of the authors and confirmed by informal listening among all authors.

2.4. Subjects

Twenty-one paid volunteers at New York University participated in this study. The order of mediations presented to each subject was chosen randomly. Three subjects were presented with the order Avatar-Static-Natural; three subjects with Avatar-Natural-Static; four subjects with Static-Avatar-Natural; four subjects with Static-Natural-Avatar; three subjects with Natural-Avatar-Static; and three subjects with Natural-Static-Avatar. The procedure took approximately 2 hours to complete, including training and testing.

3. RESULTS

Results were collected and analyzed for the training and testing phases of the experiment. We focus our analysis on the effects of training, and on the type of mediation by considering three measures of performance: localization accuracy, search time, and search strategy.

3.1. Asymptotic performance

The goal of the training phase of the experiment was to familiarize subjects with the task and interface. Training continued until the subject exhibited as asymptote in their search times; e.g. when the search times of the latest 4 consecutive trials had a standard deviation of 2.5 seconds or less.

Figure 2 shows an example of a subject's search times for all trials during training, for the three experimental conditions: static mediation (top panel), avatar mediation (middle panel) and natural mediation (bottom panel). The unfilled bar represents the trial at which asymptotic performance was reached. The subject in this example was presented with the Natural mediation first, followed by the Avatar and Static mediations. In this typical example, the subject reached optimal performance in the Static mediation at trial 10, in the Avatar mediation at trial 19, and in the Natural mediation at trial 17.

The boxplot in Figure 3 shows the results of the number of trials it took to reach asymptotic performance, for all subjects. An analysis of the means of the number of training trials for each mediation shows that it took on average 13.8 trials to reach optimal performance for the Static mediation, 19.4 for the Avatar mediation, and 20.2 for the Natural mediation.

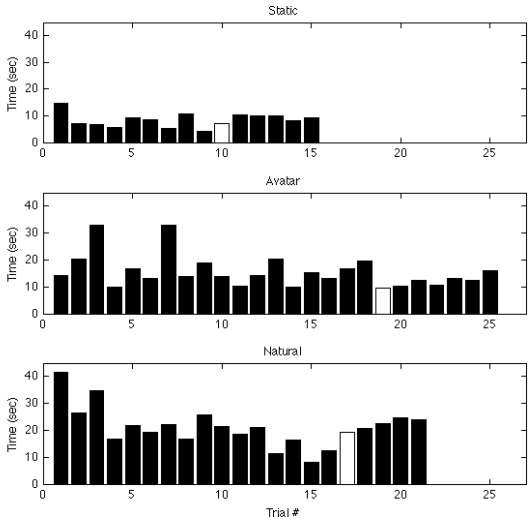

Figure 2: Example of training time results for a subject for the Static (top), Avatar (middle), and Natural (bottom) mediations. The unfilled bar represents trial when asymptotic performance was reached.

Results show that the training period was generally longer for the Avatar and Natural mediations as compared to the Static mediation. However, there was no significant difference between the training periods for the Avatar and Natural mediations.

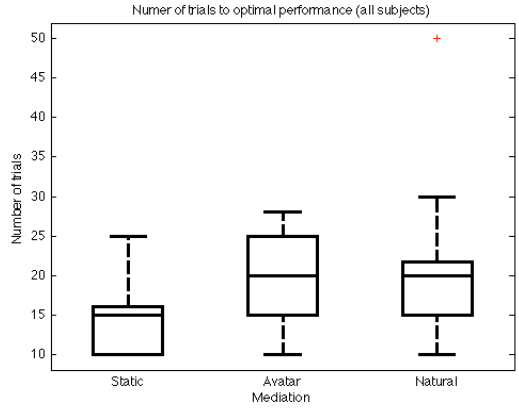

Figure 3: Number of trials to asymptotic performance for static, avatar and natural mediations.

3.2. Search time

The average search times for the training and testing phases for the three mediations are presented in Table 1. For the training phase, the search time shown represents the average of the search times once asymptotic performance has been reached. The search time presented for the testing phase is the average of all trials.

Results show that the average search time during the testing phase is 5.7 seconds for the Static mediation; 12.3 seconds for the Avatar; and 10.4 seconds for the Natural mediation. Once asymptotic performance has been reached, the search times vary little within and between subjects. Search times for the testing phase are very similar to that of the training phase – 5.8seconds for Static, 12.7 seconds for Avatar, and 10.7 seconds for Natural, representing a difference of at most 3.25% between the two phases of the experiment. A boxplot of the results is shown in Figure 4, which compares performance between the training phase (black), and testing phase (grey).

In general, Static mediation results in a much shorter search time than the other mediations, with the Avatar mediation showing slowest response time.

	Static	Avatar	Natural
Training	5.7sec	12.3sec	10.4sec
Testing	5.8sec	12.7sec	10.7sec

Table 1: Search times for training and testing phases of the experiment. Optimal search times are shown for the training phase, while mean search times are shown for the testing phase.

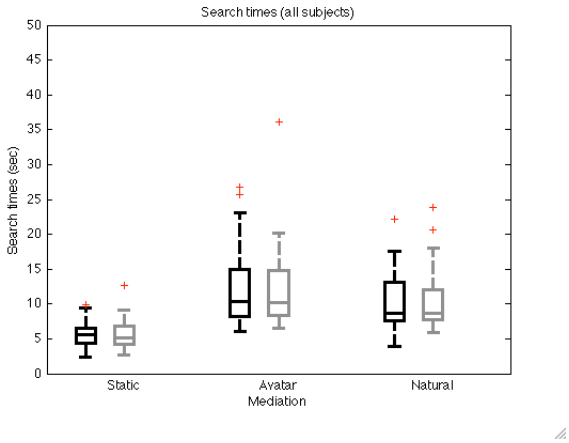

Figure 4: Search times for the static, avatar and natural mediations during the training (black) and testing (grey) phases of the experiment. Crosses represent the outliers.

3.3. Front/back confusions

One of the significant issues in localization, especially in non-interactive environments, is front-back confusion. These confusions on the median plane occur when a listener perceives a source as coming from the front, when it is in fact originating from the back (and vice versa), on the same cone of confusion.

Past research has shown that front-back confusions occur more frequently in non-interactive environments, than environments where the listener is able to rotate their head and interact with the auditory environment [5][6].

Results from this experiment show that the greatest number of front/back confusions occur in the static condition – confusions were experienced in 37.6% of all trials in the Static mediation during training, and in 32.9% of trials during testing. In comparison, only 7.7% of trials experienced front/back reversals in the Avatar mediation during testing (8.8% during training), and 8.5% in the Natural mediation (11.2% during training), see Table 2.

	Static	Avatar	Natural
Training	37.6%	8.8%	11.2%
Testing	32.9%	7.7%	8.5%

Table 2: Percentage of front/back and back/front confusions for each mediation for training and testing phases.

	Static	Avatar	Natural
Training	37.6%	30.6%	38.3%
Testing	53.5%	44.1%	47.4%

Table 3: Percentage of front/back confusions

A closer look into the specific direction reveals that the majority of confusions were front/back reversals (not back/front reversals), particularly during training (see Table 3). During the training phase, out of all the front/back and back/front confusions, front/back reversals were experienced for 37.6% of reversals in the Static mediation, 30.6% for Avatar, 38.3% for the Natural mediation. The rate of the front/back reversals increased during the testing phase to 53.5% for Static mediation, 44.1% for Avatar and 47.4% for Natural mediation.

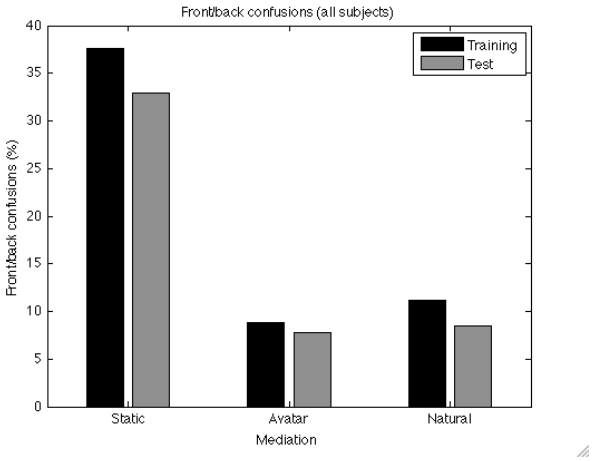

Figure 5: Percentage of front/back confusions in the static, avatar and natural mediations during training (black) and testing (grey) phases of the experiment.

3.4. Mirror localization

An effect was observed, particularly in the Avatar mediation, where subjects identified the location of the target as being 180˚ on the opposite side of the localization plane. For example, if a target source was presented at +90˚, a subject might perceive the source as originating from -90˚. We term these “mirror localizations”.

This phenomenon was observed in the Static and Natural mediations, but with a small rate of occurrence (3.5% in the Static mediation and 1.3% in the Natural mediation). It is in the Avatar mediation that we see this very prominent effect, occurring in 22.9% and 15.3% of trials during the training and testing phases, respectively.

We attribute the confusion between "left" and "right" to the additional task load imposed on the user by the visual display. We hypothesize that as participants rotate the perspective of the Avatar, affecting the yaw of the avatar, they must likewise mentally rotate their position in the environment based on their visual orientation of the avatar's nose. This rotation is not a simple one to make. We see an improvement between the training and testing phases of the experiment (with a drop in mirror localization of over 7%), but the effect remains significant.

	Static	Avatar	Natural
Training	2.4%	22.9%	2.1%
Testing	3.5%	15.3%	1.3%

Table 4: Percentage of mirror localizations during training and testing. Percentage is calculated as a function of the total number of trials for the mediation.

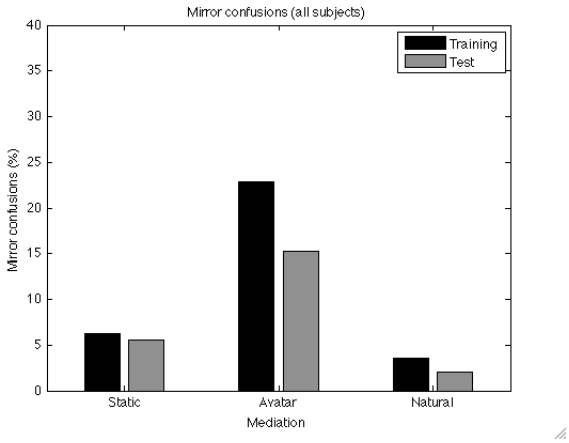

Figure 6: Percentage of mirror localizations during the training and testing phases of the experiment.

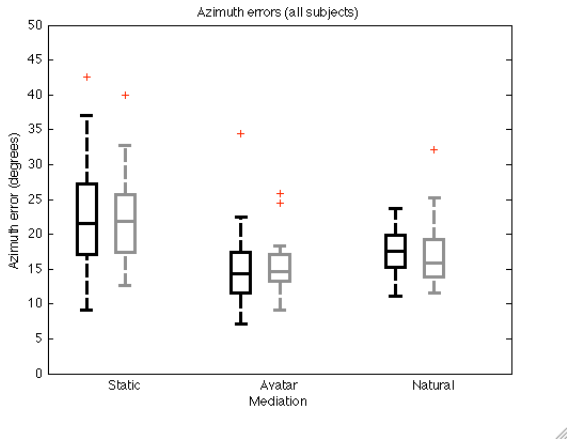

Figure 7: Boxplot of azimuth localization errors for all subjects for the Static, Avatar and Natural mediations during the training (black) and testing (grey) phase of the experiment. Crosses indicate outliers.

3.5. Localization accuracy

An analysis of the localization accuracy was performed, comparing the source location subjects identified with the actual source location. In the analysis, the locations were compensated for found front/back confusions, by reflecting across the frontal plane into the same hemisphere as the location of the target signal. To study localization accuracy, results were also compensated for mirror localizations, by collapsing the locations of trials where mirror localizations were found to the frontal right quarter plane. For example, locations of +135˚, +45˚, -135˚ and -45˚ would all be collapsed to +45˚, and a source positioned at +135˚ whose location was identified as -40˚ would result in a localization error of -5˚. To calculate the average localization error, the absolute value of the difference between the actual location and the location identified by the subject was used.

The average localization errors are presented in Table 5. Localization errors were virtually identical during the training and testing phases of the experiment. Results show that the largest localization errors were made with the Static mediation: 23.4˚ during training and 22.4˚ during testing. The Avatar mediation resulted in the smallest average localization error, 15.3˚ for training and testing. The Natural mediation resulted in an average localization error of 18˚ during training, and 17.1˚ during testing.

Boxplots of the localization errors are presented in Figure 7. Localization errors for the training and testing phases are shown in black and grey, respectively.

	Static	Avatar	Natural
Training	23.4˚	15.3˚	18.0˚
Testing	22.4˚	15.3˚	17.1˚

Table 5: Average localization errors for the three mediations during training and testing. Results shown have been compensated for front/back and mirror localizations. See text for details.

3.6. Search strategy

The listeners were observed using a number of strategies to guide them to the target. Some listeners were observed overshooting the sound source by rotating in the direction of the sound source, letting the sound source pass through the leading ear, then to the opposite ear, then correcting their position by moving the sound source back to the center of the head (directly in front of the listener). An example of this can be seen in Figure 8. Some listeners rotated until the source was directly in front of them and the target sound passed from one ear, to being at roughly equal sound levels in both ears, directly in front of the user. An example of this is in Figure 9. In some of the search attempts, users initially rotated in the opposite direction of the sound source, then corrected their positioning to identify the target source. An example of this behavior can be seen in Figure 10. In some of the unsuccessful search attempts, the listeners rotated to the sound source, but incorrectly marked its position. This type of behavior can be seen in Figure 11.

In the fastest trials of the mouse condition, 39.68% of the listeners were observed rotating until the source was directly in front of them and the target sound passed from one ear, to being at roughly equal sound levels in both ears, directly in front of the user. Another 14.29% of the listeners were observed overshooting the sound source by rotating in the direction of the target source, letting the sound source pass through the leading ear, then to the opposite ear, then correcting their position by moving the sound source back to the center of the head (directly in front of the listener). To contrast, in the slowest trials of the mouse condition, 50.79% of the listeners were observed making sudden left/right jumps in position and 44.44% of listeners made front/back jumps in position to locate the sound source. In 47.62% of the slowest trials, the listeners initiated the trial by rotating in the direction opposite the sound source.

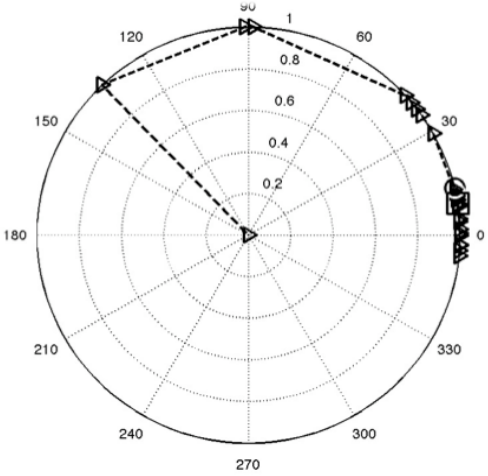

Figure 8: Example of overshooting the sound source search strategy. Triangles represent the search trail. Circle represents the true location of the target sound and the triangles square represents the location of the target sound indicated by the listener.

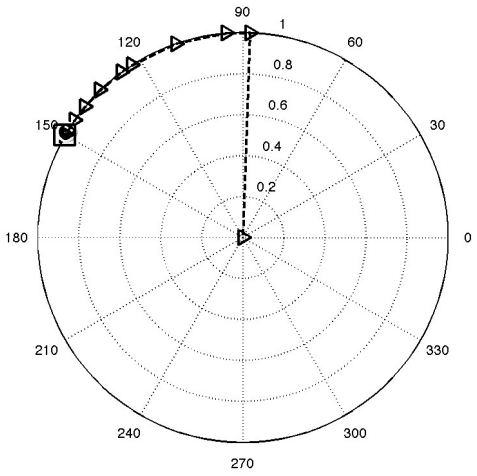

Figure 9: Example of search strategy where listener aligns the source to the frontal location. Triangles represent the search trail. The circle represents the true location of the target sound and the triangles square represents the location of the target sound indicated by the listener.

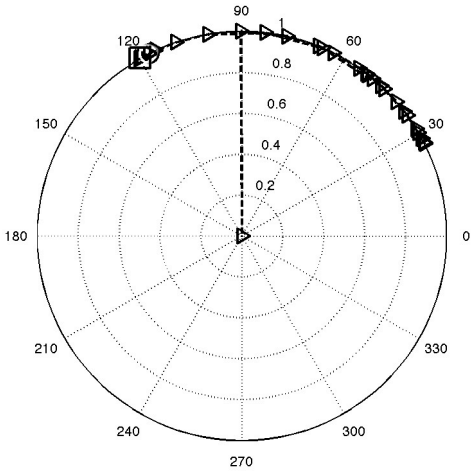

Figure 10: Example of search strategy where the listener began by rotating in the wrong direction and quickly corrected. Triangles represent the search trail. The circle represents the true location of the target sound and the triangles square represents the location of the target sound indicated by the listener.

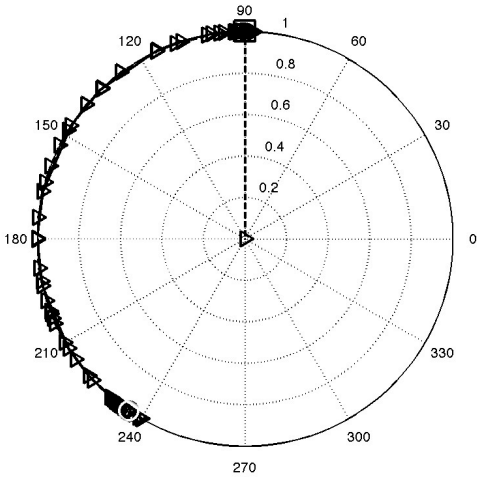

Figure 11: Example of a listener correctly rotating to the target, but indicating the incorrect location of the source. Triangles represent the beginning of the search trail. The circle represents the true location of the target sound and the triangles square represents the location of the target sound indicated by the listener.

In the fastest trials of the tracker condition, 74.60% of the listeners did not successfully mark the target source and 53.97% of the listeners rotated to the correct source location and still incorrectly identified the location of the target source. Similarly, in the slowest trials of the tracker condition, 74.60% of the listeners did not successfully mark the target source and 63.49% of the listeners rotated to the correct source location and still incorrectly identified the location of the target source. In most cases, the target source was incorrectly identified in the fastest and slowest trials for the tracker condition.

In analyzing the search strategies, for the mouse condition, the listeners were most successful using the strategy of rotating until achieving an equal sound level in both ears. Another successful strategy included rotating until the sound passed through both ears individually, then rotating back until the sound was perceived in both ears simultaneously at equal levels. The tracker condition presented much more of a challenge. Even in the fastest trials, the listeners had great difficulty locating the source, oftentimes passing through the source and still incorrectly identifying its location.

4. CONCLUSIONS

This paper presented the results of a study that compares listener performance in a navigation and search task under a static condition, and two dynamic conditions where an avatar and natural interface was used to explore the virtual auditory environment by interacting with the environment through head orientation.

In general, the results reveal tradeoffs between and within the dynamic and static interfaces. Tradeoffs exist in the length of the training period, search time of the target source, localization accuracy, and localization confusions. These tradeoffs are in addition to those inherent to the interface itself , which include environment sensitivity, motion accuracy and limitations, hardware cost, calibration, etc. While there are several advantages to the Static mode, results suggest that listeners benefit significantly from changes in orientation in a VAE, whether such changes are mediated by natural means (head rotation) or by changing the orientation of an avatar on a display.

Results from the training phases of the experiment suggest that subjects required approximately the same amount of training to reach asymptotic performance in both the Avatar and Natural mediations (19.4 and 20.2 trials, respectively). Static listening required a much shorter training period (13.8 trials). In addition to the significantly shorter training period, the Static mode also resulted in much shorter response times than the dynamic mediations. Response times averaged 5.8sec in the Static mode in comparison to 12.7 sec and 10.7 in the Avatar and Natural mediations.

Front-back confusions are reduced, on average, by a factor of 4, although some residual level of front-back confusability remains (8%). Whether this could be further reduced by eliminating search time as a variable subjects were told to minimize remains to be seen. That is, with no weight given to search time, it may be that front-back confusion can be eliminated altogether by searching for a long enough time. The data suggest that listeners don't expect to learn much from a static display.

Of the three, the static interface yielded the fastest search times, while listeners chose to spend twice as long with either dynamic interface before making their response. Of the two dynamic interfaces, an unexpected outcome was the high rate of mirror reversals found for avatar mediation. We speculate that the source of this confusion lies more in the listener's interpretation of the visual display, as opposed to a dramatic loss of sensitivity to left-right differences in the auditory display. Indeed, the birds-eye-view perspective requires the listener to reverse the "left" and "right" orientation when the head has rotated from "pointing up" to "pointing down". When coupled with a time-critical task and relatively little training, listeners may simply be misinterpreting their orientation within the VAE because of misleading display cues. Further improvements in the visual display, coupled with training, may likely eliminate this one significant difference between natural and avatar mediation of orientation.

5. REFERENCES

[1] Roginska, A., Wakefield, G.H., Santoro, T.S. McMullen, K. (2010). "Effects of interface type on navigation in a virtual spatial auditory environment". Proceedings of the International Conference on Auditory Displays (ICAD), Washington, DC.

[2] Roginska, A., Wakefield, G.H., Santoro, T.S. (2010). "Use of Interfaces in an Auditory Environment during a Navigation and Search Task". Proceedings of the Undersea Human System Integration Symposium, Providence, RI, July 27-29, 2010.

[3] Cheng, C. and Wakefield, G. H. (2001). "Moving Sound Source Synthesis for Binaural Electro-acoustic Music Using Interpolated Head-Related Transfer Functions (HRTF's)," Computer Music Journal, 25(4), 57-80.

[4] The BBC Sounds Effects Library. Princeton, N.J.: Films for the Humanities & Sciences vol. 1-40 (1991)

[5] Wallach, H. (1940). "The role of head movements and vestibular and visual cues in sound localization," Journal of Experimental Psychology, vol. 27, 339–368.

[6] Wightman, F.L., and Kistler, D.J. (1999), "Resolution of front-back ambiguity in spatial hearing by listener and source movement," Journal of the Acoustical Society of America, vol. 105, 2841–2853.