

An Optimized Spatial Audio System for Virtual Training Simulations: Design and Evaluation

David L. Jones, Kay M. Stanney

Human-System Integration Lab,
College of Engineering and Computer Science,
University of Central Florida,
FL, USA
david, stanney@mail.ucf.edu

HeshamFoad

VR Sonic, Inc.
Arlington, VA, USA

hfouad@vrsonic.com

ABSTRACT

This paper explains the design and evaluation of an optimized spatial audio system for a close quarters combat virtual training system. The study explained herein compared performance levels of participants training with the system based on the auditory spatialization fidelity level that sound was presented at. The three levels of fidelity that were evaluated were nonspatialized audio, and the use of a generalized HRTF model, or best-fit HRTF models to present audio. This study focused on the capability of spatialized audio to direct a search and detection task in an applied training environment. The results demonstrate that the use of spatialized audio systems lead to better performance on applied search and detection tasks over nonspatialized audio or systems lacking sound altogether. On the other hand, the results do not show performance differences between participants using generalized HRTF's when compared to best-fit HRTF's. The results help to extend studies evaluating the use of 3D audio systems [1,2,3,4,5] to a more applied setting and may be of interest to those designing virtual training systems with an integrated audio component.

1. INTRODUCTION

Virtual Reality (VR) training systems provide a safe means to train for tasks that would otherwise be dangerous or impossible in the real world. This reason coupled with the decreasing cost of VR systems is leading to an increased use of them to train people on various tasks, ranging from complex surgical procedures to military operations. As these systems are becoming more common in the training domain, it is important to evaluate the effects of auditory displays on these virtual training environments.

Research by Mulgund et al. [1], suggests that spatialized audio can be used to communicate direction, location, movement, and aid in guiding navigation. Research by Makino et al. [2] supports this idea that spatialized audio can lead to faster and more efficient navigation. Furthermore, by augmenting visual displays with spatial audio, time required to perform target detection tasks have been reduced when targets are in the within the current field of view [3], as well as when audio is used to direct attention to targets in the peripheral visual area [4]. In addition to target localization, Rudmann et al. [5] demonstrated that spatial audio may be effective in aiding target identification as well.

The results of the research listed above illustrate how useful the integration of spatialized audio within visual systems can be. It is important to apply this research to virtual training

environments in order to evaluate the effects of spatialized audio when participants are performing a task less trivial than a target detection task. In order to do this, a spatial audio presentation system was designed to present audio at 3 spatialization fidelity levels while users were training to perform a Close Quarters Battle for Military Operations in Urban Terrain (CQB for MOUT) room clearing task. Based on the research described above, the inclusion of spatialized audio was expected to lead to better performance when detecting and engaging enemy and friendly units in the virtual environment.

2. SYSTEM DESIGN

The auditory component of the training system was developed using a software product called ViBeStation. ViBeStation is an auditory scene editing and runtime tool that supports distributed operation in a simulation environment. ViBeStation includes a number of features that facilitated the development of the audio scenarios for VE training:

- An auditory scene editing capability facilitated the creation of training scenarios.
- Configurable auditory displays allowed the rendering of the auditory scene at varying levels of fidelity
- An auditory scene storage and retrieval capability
- Distributed operation using the HLA RTI simulation infrastructure.
- An event logging capability for capturing simulation events for performance analysis
- A best-fit HRTF selection capability

The experimental setup consisted of two dual-processor Dell Precision computers. One computer was used to produce stereo graphics and the other produced spatialized audio cues using ViBeStation software. Participants were immersed in the VE using a Virtual Research Systems V8 Head Mounted Display (HMD). Audio cues were presented using Sennheiser headphones. Head tracking was done using an Intersense InertiaCube tracker and was used to control the point of view and direction of motion in the VE. A Saitek P2500 Rumble Pad controller was used to control movement, firing, and friendly clearing input by the participant.

2.1. Auditory Scene Analysis

An auditory scene analysis was carried out in order to determine important auditory cues for the CQB MOUT task of room

clearing. The scene analysis was based on Greenwald’s Critical Cue Inventory [6], which was applied to auditory cues associated with the room clearing task. An auditory cue strategy was developed based on observational analysis of a room clearing exercise. In addition, interviews with subject matter experts (SMEs) were used to clarify work processes and practices associated with room clearing exercises and determine where auditory cues should be focused. The resulting strategy included the use of a number of cues from Greenwald’s CCI that are naturally present in the real-world environment, as well as metaphoric cues that are expected to train participants to avoid dangerous situations. The naturally present sounds were targeted at assisting Virtual Environment (VE) trainees in locating and differentiating between hostile and friendly units and include enemy voices, movements, and weapon sounds, as well as the movement and voices of friendly units present. Additionally, the sounds of shots fired are played from the position of each hostile unit located in a room whenever the trainee entered that room.

The metaphoric cues were aimed at assisting trainees in developing effective strategies during training sessions. They included earcons or auditory icons that were played whenever participants were in dangerous areas, such as in doorways, room entry areas, or in front of windows or mouseholes. The cue used to symbolize when participants were in a doorway, room entry area, or in front of a mousehole was the repeating resonance of a foghorn. Whenever participants were standing in front of windows, an auditory icon resembling the sound of breaking glass was played. Finally, whenever participants were too close to walls when walking down a hallway, the sound of a pan flute playing a middle C quarter-note was repeated to create a warning cue. By cueing trainees that they are in such areas, they could change their room clearing strategy in order to avoid such obstacles. Essentially, the metaphoric cues provide during-action feedback to help train participants on what they should do to minimize the threat to themselves while performing MOUT CQB tasks. The cues that were integrated into the VE and the information that they were expected to provide are presented in Table 1 below.

Cue Category	Integrated Cues	Information Provided
Naturally present	Enemy voices	Enemy presence and position
	Enemy movement	
	Weapon sounds	
	Non-hostile voices	Non-hostile unit presence and position
Non-hostile movement		
Metaphoric	Window danger area auditory icon (shattered glass)	Alarm to signal that trainee is silhouetting in front of a window
	Entryway danger area earcon	Alarm to signal that trainee is standing in front of an open door of an uncleared room
	Mousehole danger area earcon	Alarm to signal that trainee is standing in front of a mousehole
	Wall proximity earcon	Alarm to signal that trainee is standing too close to a hallway wall

Table 1: Integrated cues and provided information

3. METHOD

3.1. Participants

Thirty-six university students (26M; 10F), with an average age of 23.8, ranging from the ages of 18 - 49 participated in this study. On average, participants spent 27.5 hours (+/- 18.42) per week on the computer, and 8.07 (+/- 7.42) hours of that time playing video games.

3.2. Audio Conditions

Three audio spatialization fidelity levels were evaluated throughout the study. Additionally, a no sound condition served as the control group. Participants of the control group did not receive any auditory feedback during any trials. Participants in the nonspatial audio condition heard both environmental and training (during training sessions) sound cues while performing the room clearing task. The sound cues in this condition were occluded based on the position of walls but did not have a directional component when heard by the participant. The second level of spatialization fidelity that was evaluated was the generalized HRTF model. The sound cues in this condition were spatialized with the generalized HRTF model using an HRTF dataset obtained from a KEMAR dummy-head microphone. The final level of audio spatialization fidelity evaluated was the best-fit HRTF condition. A two step subjective analysis outlined in [7] was used to select the best-fit HRTF for each participant in this condition. Before running any trials, participants were required to perform a subjective analysis of the 45 CIPIC HRTFs. During the first step participants selected 5 HRTFs out of the possible 45 HRTFs based on the best spatial perception in the frontal area. The 5 HRTFs were then compared in a pair-wise method based on how well audio was spatialized 360 degrees around them. To do so, they were allowed to listen to a white noise signal presented at 45 degree increments around them for each of the top five models that were selected. The selection criteria for the second step were as follows:

- Sounds moved horizontally in equally-spaced steps
- Sounds maintained a constant elevation at all times
- Sounds were perceived in the correct area (front or back)
- Sounds were perceived at a constant distance outside the head

The comparisons of the top five models were used to select the model that best-fit that participant. The selected model was then used for presentation of all sound (training cues and environmental sounds) to participants.

3.3. Tasks

Each participant performed a series of CQB for MOUT room clearing tasks in a virtual environment. The primary objective of participants was to clear each environment that they were immersed in as efficiently as possible. This task required that participants move down a hallway, enter and clear all open rooms and engage all hostile and non-hostile units located therein. Whenever an enemy or nonhostile unit was detected, the participant was required to shoot or press a button to acknowledge that they are not a threat. At no time during the training were participants told to hold either of these tasks (firing at enemies or acknowledging non-hostiles) at a higher priority than the other. Performance on this task was based on

the time required to clear friendly units, time required to fire on enemy units, and the total number of hostile and non-hostile units engaged.

While performing the task above, participants were also required to avoid any areas in the environment that would pose a threat to them. During a training session and before performing in the VE participants were instructed to avoid standing in front of open doors or holes in the walls while in the hallway, in the entrance area directly inside of rooms, in front of windows, or within 6" of any wall while moving down the hallway.

3.4. Procedure

Before the start of the test session, participants completed an informed consent, demographics questionnaire, and a simulator sickness questionnaire to evaluate pre-exposure levels of sickness. Once these forms were completed, participants in the Best-fit HRTF group performed the HRTF selection procedure described above. All participants then viewed a presentation that trained them on how to interact with the virtual environment that they were to be evaluated in. This training instructed participants on how to move and engage enemy units in the Virtual Environment, how to differentiate hostile from non-hostile units, the rules of engagement for the exercise, and the location of danger areas that they were expected to avoid. After this training session, each participant was immersed in a 7 minute virtual training scenario where they were guided by the experimenter to familiarize them with how to move and engage enemy and friendly units in the VE. Following this initial training scenario, participants completed two 5 minute familiarization scenarios in a VE that closely emulated the environment that they later would be evaluated and trained in. None of the familiarization scenarios included the use of any sound cues.

After the three system familiarization scenarios were completed, each participant performed the building clearing task in a 15 room environment (see figure 1 for room layout) under the sound condition of the group that they were randomly assigned to. After completion of this trial, participants filled out a workload and situational awareness questionnaire.

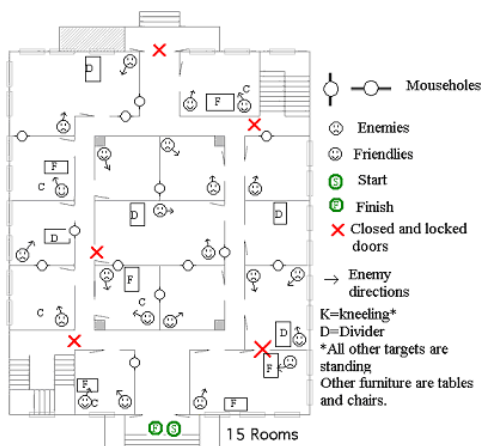


Figure 1: Pre and Post Training Evaluation Floorplan

Participants were then given additional training on the location of danger areas in each environment and how to use the audio presented in the scenario to more efficiently perform the

task. Each participant then performed the building clearing task in 4 separate 8 room environments (see figure 2 for example layout). For the participants in each sound condition, the system played the metaphoric cues to indicate when participants were positioned in the danger areas described above. In addition to these cues, all of the natural environmental sounds that were present in the initial trial were also presented to the participants in each audio condition. Once the four training environments were complete, participants filled out a workload and situational awareness questionnaire.

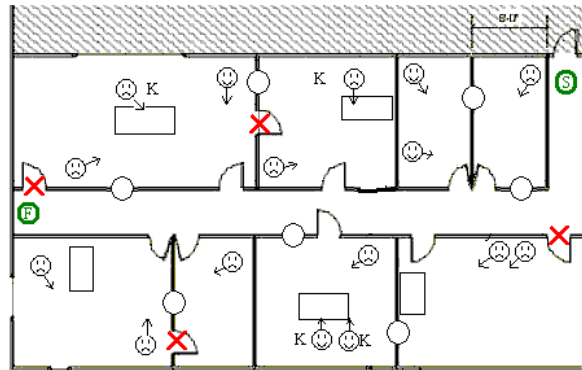


Figure 2: Example Training Scenario Floor Plan

After completion of the questionnaires, the participants were evaluated in a final trial. During this trial, each participant was required to perform the building clearing task in the same 15 room environment that the initial trial was performed in. This trial was identical to the initial evaluation trial. During this final trial, only natural environmental sounds were presented to the participants in the three sound conditions. Following the completion of this final trial, participants filled out a workload, situational awareness, presence, and overall user reaction questionnaires.

Once all trials and evaluations were complete, participants in each of the spatialized sound conditions performed an auditory spatialization test in order to evaluate the localization acuity of the participant. In this test, participants were seated in a stationary position surrounded on both sides and the front with fabric material to prevent them from seeing the room outside of the enclosure. Participants were directed to look directly forward at an X marked on the fabric in front of them while bursts of 499 Hz sine waves were played from various locations on a 4 foot radius circle surrounding them. After each sound was played, participants were required to mark the position of the sound source on a diagram. This evaluation was later used to determine the participant's audio localization abilities.

3.5. Experimental Design

In order to evaluate the effects of various audio fidelity conditions on performance, a mixed 4x2 (audio condition x training level) mixed design was used. The between-group factor evaluated is the type of sound condition that the participant was assigned to. Performance measures were taken both pre and post training and were evaluated as the within group factor. ANOVAs were performed on the data collected pre-training and post-training.

4. RESULTS

Performance measures including average time spent clearing each room and average time required to engage enemy units were compared between the sound conditions in order to evaluate the value of spatialized sound for detecting and engaging hostile units. It is expected that if spatialized audio is a useful cue to direct users to dangerous areas, the spatialized audio conditions will outperform non-spatialized and no difference should be found between the non-spatialized and no audio conditions.

Although the average time spent clearing each room did not show a significant difference ($\alpha = .05$) based on audio conditions ($F(3, 32) = 2.41, p = .085$), a strong trend was present suggesting that the average time required to clear rooms decreased when spatial audio was used when compared to nonspatialized audio (see figure 3)

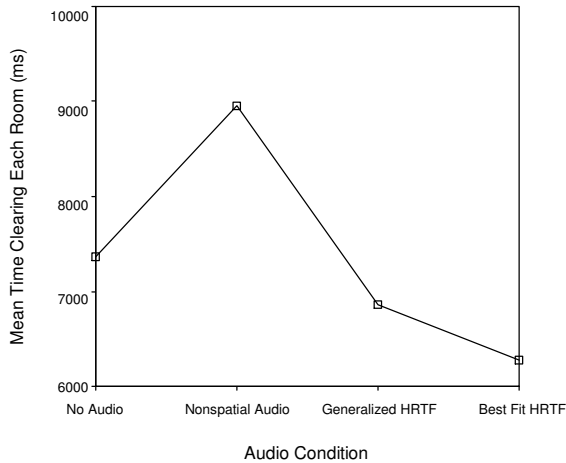


Figure 3: Mean Times Clearing Room for Each Condition: Pre-training

Once participants were more experienced at the task, the trend of decreasing time required to clear each room as the audio spatialization fidelity increased became more apparent. This trend can be seen in Figure 4

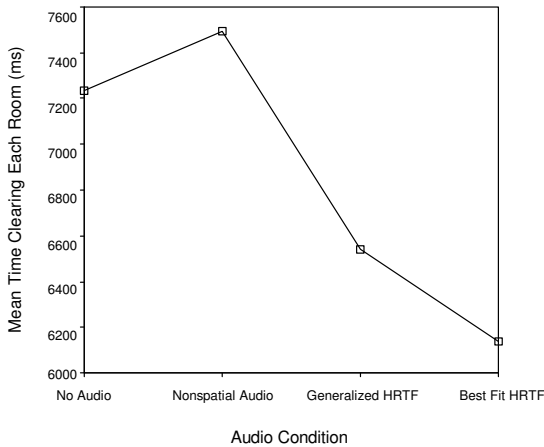


Figure 4: Mean Times Clearing Room for Each Condition: Post-training

The average time required to engage enemy units after entering rooms although not significant ($F(3, 32) = 2.301, p = .096$), showed a strong trend suggesting that when participants were untrained at using the audio to direct the search for hostile units, spatialized audio systems led to shorter search times. This trend is apparent in figure 5 below

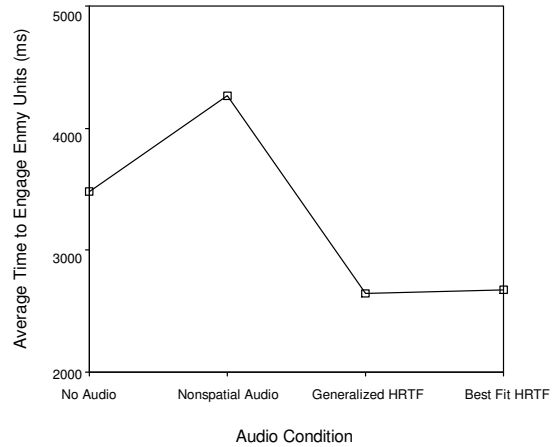


Figure 5: Mean Times Engaging Enemies After Room Entry: Pre-training

After participants had more time to train with the system and learn how to use the spatial audio to detect enemies within the environment the same pattern was apparent at a significant level ($F(3, 32) = 3.02, p = .044$) as can be seen in figure 6.

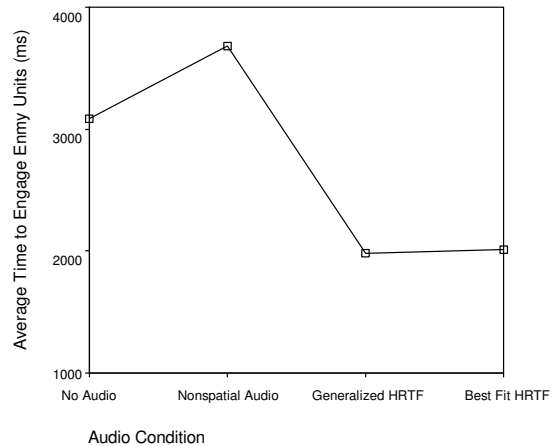


Figure 6: Mean Times Engaging Enemies After Room Entry: Post-training

A post-hoc analysis of the average time required to engage enemy units after entering a room showed that the use of nonspatial audio led to longer search times for enemy units over both generalized and best fit HRTF systems ($p = .018$ and $p = .020$, respectively). There were no significant differences found between the use of no audio and nonspatialized audio or between generalized and best fit HRTF systems.

5. CONCLUSIONS

The results found in this study support the hypothesis that spatialized audio systems will lead to more efficient room clearing. For each measure of efficiency, participants using the spatialized audio system outperformed participants using no audio and nonspatialized audio systems. This suggests that spatialized audio can be used to direct search and detection patterns in an applied setting.

It is important to note the increased significance of the differences between the groups after participants had the opportunity to train with the different audio cues. This suggests that although novice users will get value out of using a spatialized audio system (generalized and best fit) for this search and detection task, the advantages will be more apparent after they gain some experience training with the system.

The lack of performance differences between the generalized HRTF displays and best-fit HRTF displays suggests one of two things. It is possible for this real world search and detection task, generalized HRTFs are just as useful as best-fit HRTF's. It is also possible that the best-fit selection process that was used did not lead to participants choosing the best HRTF model for them. If this is the case, the subjective best-fit HRTF selection process should be reevaluated and compared to more objective means of selecting best fit HRTF's. This will assure that the lack of significant performance differences isn't due to the selection process used.

The results of this study support the use of spatialized audio in Virtual Reality training systems. In this case, it provides participants with a cue naturally available in the real-world that can be used to direct how they interact with the environment. This leads to a training system that more closely matches the real world environment and is expected to increase the transfer of training into such environments.

6. REFERENCES

- [1] Mulgund, S., Stokes, J., Turieo, M. and Devine, M. (2002, October). *Human/Machine Interface Modalities for Soldier Systems Technologies (Report No. 71950-00)*. Cambridge, MA: TIAZ LLC. (DTIC No. ADA414918)
- [2] Makino, H., Ishii, I., Nakashizuka, M. (1996). Development of navigation system for the blind using GPS and mobile phone combination Proceedings of 18th Conference IEEE, EMBS, vol. 18 1996 pp. 1-5.
- [3] Bolia, R. S., D'Angelo, W. R., & McKinley, R. L. (1999). Aurally aided visual search in three-dimensional space. *Human Factors*, 41(4), 664-669.
- [4] Perrott, D.R., Saberi, K., Brown, K., & Strybel, T.Z. (1990). Auditory psychomotor coordination and visual search performance. *Perception and Psychophysics*, 48, 214-226.
- [5] Rudmann, D. S., & Strybel, T. Z. (1999). Auditory spatial facilitation of visual search performance: Effect of cue precision and distractor density. *Human Factors*, 41, 146-160.
- [6] Greenwald, T. (2002). An Analysis of Auditory Cues for Inclusion in a Close Quarters Battle Room Clearing Operation, MOVES Institute, Naval Postgraduate School, Monterey, CA.
- [7] Seeber, B.U., Fastl, H. (2003). *Subjective Selection of Non-Individual Head-Related Transfer Function*, Proc. 2003 International Conference on Auditory Display, pp. 259-262, Boston University, Bost, MA, July 6-9, 2003.