SPATIAL PERCEPTION AND EXPECTATION: FACTORS IN ACOUSTICAL AWARENESS FOR MOUT TRAINING

Darin E. Hughes, Jennifer Thropp, John Holmquist, J. Michael Moshell
University of Central Florida
Orlando, FL 32816

ABSTRACT

Mixed Reality (MR), and its predecessor Virtual Reality (VR), has been primarily viewed as a visual science, with much less attention given to the other senses, despite clear evidence of their importance, especially audio. In fact, in military operation in urban environments, audio is often a more primary sense than vision, providing a soldier with an early warning system that needs to be honed and trained. Our research program, in contrast, treats the auditory sense as an equal to the visual. The consequences are MR experiences that have much greater impact than those in which audio is just an after thought. However, given the depth and breadth of graphics research, we are compelled to learn from this mature area. Thus, we are constantly striving to find results from graphics research that have useful analogies in the audio domain. Lessons learned from these analogies, especially as concern people's perception and expectations, are the focus of this paper.

1. SUMMARY

Graphics research is essentially about algorithms and a never-ending search for faster ways to render complex scenes. Faster graphics hardware, in the form of programmable graphics units (GPUs), have made it possible to achieve results that only a few years ago seemed unachievable on consumer grade machines. Nonetheless, the search for faster algorithms never ends, as user expectations rise to meet and exceed advances in hardware. To keep up with such demands, graphics researchers constantly take advantage of limitations in human perception and the expectations of viewers. The simplest example of gaining performance from human limitations is the use of clipping planes and object culling to reduce what must be rendered; after all, people cannot see behind their heads, and their peripheral vision, while acutely aware of motion, does not discern details well. As regards expectations, all computer graphics are based on the user's belief system, since what can be represented is so much more restricted than what can actually be discerned. Even simple cell animation plays on our expectations of continuity. What then can be learned from this that can lead to a less costly, more effective soundscape? This paper addresses the topic of expectation in audio, and how people's experiences play such a strong role in their audio belief systems.

Despite attempts in software and hardware to deliver three dimensional audio, there has been little research on the aesthetic effects of sound design, and the influence of expectation in spatial perception and other more subjective measures (Begault, 1999). Our experiments indicate that expectations play a crucial role in our perception of sound localization, a key skill that soldiers use in identifying danger. In this context, expectations refer to the extent to which a sound is associated with a particular location (Cheung, 2002). In a pilot study conducted with 21 participants, sounds such as airplanes, helicopters, and lightning were perceived as being above head-level even when the sounds were played from speakers positioned only at head-level. Additionally, sounds such as a voice saying "hi" and ocean waves crashing were perceived by the participants as coming from in front of them even though the sounds were, in fact, played from four speakers in front and behind them.

The potential implications of our study may suggest that expectations are more significant than physical cues in our perception of spatialized sounds. The knowledge gained by these experiments suggests a new paradigm in spatial audio research and sound design, and in training soldiers to avoid the miscalculations that can come with an over-dependence on expectation.

2. INTRODUCTION

A human's audio perception is the only sense that can operate in 360 degrees simultaneously and provide cues around corners and through walls. Sound can also be used as a means of depth perception (Loomis, Klazky, & Golledge, 1999; Wightman & Jenison, 1995). In combat, a soldier is able to articulate acoustical characteristics and spatial registration instinctively. With proper training, soldiers use their audio sense to gain a tactical advantage. Sound cues are used to inform the listener and diagnose problems (Gaver, Smith, & O'Shea, 1991). Within realistic military operations, there are significant distractions and emotional impacts in the audio landscape that are not represented in current live or virtual simulated training, yet can have significant influence on a soldier's performance. When response time is measured in

seconds, one missed cue can be a matter of life or death. Audio systems for Military Operations in Urban Terrain (MOUT) simulations require a robust design capable of capturing, synthesizing and delivering a realistic scenario with natural cues and immersive ambience, in addition to providing synthesized cues for effective communication.

With Mixed Reality, a training scenario is able to blend the realistic impact of live simulation with the dynamic control of a virtual reality environment. MR represents the entire spectrum of simulation ranging from live to virtual, including Augmented Reality (putting virtual entities within the real world) and Augmented Virtuality (placing real entities within a virtual world). The core research goal of the MR Testbed, the context in which this research was done, is to melt the boundaries between what is real and what is synthesized, and to blend the two seamlessly together for training, entertainment and educational applications.

The first stage of our Mixed Reality Audio Research initiative is being demonstrated within the Mixed Reality in Military Operations in Urban Terrain (MR MOUT) project (figure 1). It includes the creation of core capabilities and their integration into a richly layered, multi-modal audio and visual landscape of virtual and real components (Hughes et al., 2004). Going beyond the basic mechanics of spatial reproduction, our research is looking at the psychological and perceptual factors of sound design and their potential impact on training simulations.

Graphics have long taken advantage of perceptual rendering to mask problems or to simplify complex animations (Harrison, et al., 2004). Techniques of perceptual rendering in graphics include issues related to field of view, depth perception, and color perception. Specifically, studies have been shown that expectations play a factor in the perception of direction in motion (Sekuler, Watamaniuk, & Blake 2002). This led us to believe that there may be similar expectations that influence spatial audio perception.

Figure 1. MR MOUT set

The basis of this research is grounded in the fact that audio cues serve as information-carrying channels (Gaver, Smith, & O'Shea, 1991). Sounds help people to diagnose problems and to monitor events within their environments. For instance, the repetitive mechanical sounds created by machinery indicate that the equipment continues to function at a steady rate. Conversely, a sudden cessation of this sound would suggest that the machinery has discontinued its activity, perhaps due to a malfunction. The human listener in this situation would then direct his or her attention to the appropriate portion of the environment to resolve the problem. By replicating real world environments in a virtual sense, training opportunities flourish. The person who listens for audio cues in virtual environments (VE) will have a distinct advantage, which helps to promote the feeling of presence.

One way to increase the effectiveness of VEs is through improving the user's feeling of presence. Sense of presence is the perceptual illusion of existing within the VE (Lombard & Ditton, 1997; Slater, Usoh, & Steed, 1994) and is based on stimulation of sensory, cognitive, and affective processes (Cheung, 2002). This continuous stimulation can be presented to the user in both visual and auditory modalities. Developers have strived to maximize perceived realism in a number of ways, primarily by focusing upon the visual domain. However, comparatively less research has focused upon ways to improve the realistic and accurate representation of the audio stimuli used. Yet, it is important to pair visual cues with audio cues because multimodal stimulation can increase presence (Gilkey & Weisenberger, 1995; Larsson, Västfjäll, & Kleiner, 2001).

However, there are other ways to enhance the feeling of presence. By understanding the user's expectations of where real world sounds are thought to occur in a three-dimensional space, researchers can improve the feeling of presence within the auditory domain. According to (Cheung 2002), expectations refer to the extent to which a person is primed to hear a sound in a particular location. Furthermore, sounds that are congruent with expectations yield an accurate representation of the context. Thus, a user who hears sounds that are specific to a certain environment will then be able to correctly identify that environment (Cheung, 2002). Additionally, other studies have found that expectation plays an important role in enhancing a sense of place (Chueng & Mardsen, 2002; Serafin, 2004; Turner, McGregor, Turner, & Carroll, 2003).

According to the ecological approach, human listeners describe and identify their environments based upon real world sounds (Cheung, 2002). Accordingly, listeners are more oriented towards familiar sounds of everyday events, such as doors closing, human voices, and thunderstorms, rather than qualities such as pitch and

loudness. Thus, ecological sound design can also be implemented to impact the user's sense of presence and can aid in localization.

Surround loudspeakers can place a sound source either to the side or behind the listener, allowing for specific sound effects. Such an arrangement satisfies an effective speaker configuration regarding the sound motion along the horizontal (x, z axes) plane (Soulodre, Lavoie, & Norcross, 2003). On the other hand, there is less understanding of sound presentation on the y-axis in simulating motion along the vertical plane. Thus, the efficient representation of both axes is critical because a strong sense of spatial impression is important in obtaining a subjectively pleasing and realistic sound field (Soulodre, Lavoie, & Norcross, 2003).

Finally, financial considerations are of interest, as many institutions seek cost-effective VE equipment. While accurate auditory stimuli can be generated with a large number of speakers, purchasing costs and intrusion on available space would also increase accordingly. Thus, it is of interest to determine the efficiency of a speaker system which is comprised of a smaller number of speakers.

Initial studies performed in our labs indicate that expectations play a crucial role in perception of sound localization. In a recent study conducted with 21 participants, sounds such as airplanes, helicopters, and lightning were perceived as being above head-level even when they originated from speakers positioned at or below head-level. The details of this study are described below.

3. PURPOSE OF INVESTIGATION

By understanding the user's expectations of where real world sounds are thought to occur in a three-dimensional space, researchers can improve the feeling of presence within the auditory domain. In maximizing VE realism and overall sense of presence, it is beneficial to represent auditory stimuli both accurately and adequately. Our research aimed to investigate both of these facets through an experiment which tested the impact of users' expectations upon localization of real world sounds within a representation of a VE. A better understanding of users' abilities to localize such sounds will help system developers produce simulations with the fidelity necessary to result in effective transfer of training, while also considering efficiency and cost.

4. METHOD

4.1 Participants

Twenty-one participants from the University of Central Florida digital media program received course credit for completing the experiment. Each completed a short demographic survey and signed an informed consent form. Participants were all students, but the task required no special training to complete. All participants reported normal hearing function.

4.2 Stimuli

A total of ten sounds were presented to the participants. Eight of the sounds were real-world sounds, including an airplane, footsteps, a voice saying 'hi', a helicopter, a car, the ocean, a dog barking, and thunder.

The remaining two sounds were presented in the form of "pink noise". Pink noise is a variant of "white noise.". White noise is a sound that contains every frequency within the range of human hearing (generally from 20 hertz to 20 kHz) in equal amounts. Most people perceive this sound as having more high-frequency content than low, but this is not the case. This perception occurs because each successive octave has twice as many frequencies as the one preceding it. For example, from 100 Hz to 200 Hz, there are one hundred discrete frequencies. In the next octave (from 200 Hz to 400 Hz), there are two hundred frequencies. Pink noise is white noise that has been filtered to reduce the volume at each octave. This is done to compensate for the increase in the number of frequencies per octave. Each octave is reduced by 6 decibels, resulting in a noise sound wave that has equal energy at every octave.

4.3 Apparatus

Sounds were played for 5 seconds at 80db using a Windows XP based PC using Cakewalk Sonar Studio Edition 3 software for sound control. Sounds were sent directly to each speaker using discrete channels. No surround sound features or 3D/API/EAX type enhancements were used so that each sound was unaltered on its way to the designated speaker or speakers. The speakers were set up in one tier placed at head-level, specifically 5' 7" above the floor. This tier was a 4.0 design with the speakers placed in the center of the wall between the corners.

4.4 Procedure

All data were collected at the University of Central Florida in Orlando. The experimenters first briefed each participant on the correct method for reporting the various sound locations heard during the study. The self-report forms contained a separate area to mark the start and stop location of each of the 10 sound trials presented. Participants indicated perceived start and stop locations for both the horizontal plane and the vertical plane, such that there were 4 entries per trial – a horizontal plane start location, a horizontal stop location, a vertical plane start location, and a vertical stop location. Participants

selected these locations from a choice of three levels of speakers: below-head, head, and above-head level. They were then escorted into the room where the actual trials took place. Every participant stood in the same location during the experiment. This location was set in the center of a 20' x 20' room with a one tier sound system composed of four head-level speakers. This location was also the center-point of the speaker setup. Participants faced forward at all times and were instructed to keep their heads and bodies still during each trial. They held a pen and clipboard while standing during the experiment for the purpose of responding.

Participants were asked to indicate the starting and stopping locations for each of the 10 sounds that were administered. All sounds were presented at the head-level speakers. At the completion of each sound, participants marked their entries on the reporting form by placing a checkmark in the box on the form corresponding to the location where they believed the sound was originating. There was a 10 second interval between trials. After the 10 trials were over, each participant handed the reporting form to the experimenter, and then was escorted away from the area to avoid contact with other participants who had not yet completed the experiment. Trials were administered one participant at a time.

5. RESULTS

Drawing from the data shown in Table 1, accuracies between sounds with specific location expectations from above-, at-, and below-head level were compared. The results show that sounds with the expectation of originating from above (e.g., airplane, helicopter) are very effectively perceived as originating from above, with a mean percentage score of 76.17, even though all sounds were presented at head level. The perception of these above-head sounds was significantly greater than those with an expectation of head- or below-head level $F(1, 6) = 6.517$, $p < .05$. In contrast, the sounds that have the expectancy of originating from below (e.g., footsteps, dog barking) did not show a significant difference in comparison to the tested ambient sound (pink noise), $F(1, 3) = 3.04$, $p > .05$. This showed that there was no statistically significant expectancy of the location of pink noise sounds.

Once the expected values were averaged, the above-head level expectancy was the strongest (69.87%), while head-level and below-head level expectancies were only 28.6% and 34.13% accurate, respectively.

Sound	Expected location	Expected correct		Actual	
		Start	Stop	Start	Stop
Airplane	Above-head	80.9	66.7	14.3	33.3
Pink Noise	n/a	n/a	n/a	38.1	38.1
Footsteps	Below-head	23.8	19	33.3	52.4
Voice: "Hi"	At-head	33.3	52.4	33.3	52.4
Pink Noise	n/a	n/a	n/a	38.1	38.1
Helicopter	Above-head	85.7	76.2	14.3	33.3
Car	At-head	76.2	76.2	76.2	76.2
Ocean	At-head	42.9	42.9	42.9	42.9
Dog	Below-head	19	23.8	47.6	47.6
Thunder	Above-head	61.9	47.6	33.3	33.3

Table 1. Percent correct by expected and actual sounds

CONCLUSIONS AND FUTURE WORK

In observing the behavior of participants in this study, it has become clear that expectations play a crucial role in the perception of reality. For instance, the airplane, helicopter, and thunder sounds were indeed perceived as originating from above-head level, despite the fact that they were presented at head-level. Sounds that do not carry location expectations, such as pink noise, were not perceived with any particular expected location. The potential implications of this study may suggest that expectations are more significant than physical cues in our perception of everyday sounds. This knowledge may also indicate that expensive hardware and CPU-intensive 3D audio software may not be necessary for certain applications; the user's conceptualization of environmental norms may suffice in certain applications. Additionally, the recognition of this over-reliance on expectations may prove very useful for combat training and simulation in general whereby trainees may incorrectly perceive an audio cue based on expectations rather than physical cues. This could prove a deadly mistake and future simulation-based training may aim at correcting these false assumptions.

Additional work on the subject of expectation and spatial perception is planned. A more detailed list of sounds with a larger experimental group should provide a deeper understanding of this phenomenon as well as indicate which types of sounds are most likely to induce this phenomenon. Studies are also being designed to test the effectiveness of various capture techniques for increasing immersion and presence in interactive environments.

ACKNOWLEDGEMENTS

This research was conducted within the Mixed Reality Audio research test bed at the University Central Florida's Media Convergence Laboratory, a partnership of the College of Arts and Sciences, School of Computer Science, and the Institute for Simulation and Training. The research is in participation with the Research in Augmented and Virtual Environments (RAVES) supported by the Naval Research Laboratory (NRL) VR LAB. It is also supported by the U.S. Army's Science and Technology Objective (STO) Embedded Training for Dismounted Soldier (ETDS) at the Research, Development and Engineering Command (RDECOM). Collaborative research and an ongoing dialogue are also taking place with the Army Research Institute (ARI) at the Science and Technology Testing Center (STTC) in Orlando, Florida. Finally, special thanks are due to Scott Vogelpohl, who programmed the audio engine and to the Mixed Reality Laboratory, Canon Inc., whose support of our efforts through the generous provision of their Coastar Head Mounted Display has made this research possible.

REFERENCES

Begault, D. R., 1999: Auditory and Non-Auditory Factors that Potentially Influence Virtual Acoustic Imagery, *AES 16th International conference on Spatial Sound Reproduction*, Rovaniemi, Finland, April 10-12.

Cheung, P., 2002: Designing Sound Canvas: The Role of Expectation and Discrimination, *CHI Conference on Human Factors in Computing Systems*, Minneapolis, Minnesota, April 20-25, 848-849.

Chueng, P. and Marsden, P., 2002: Designing Auditory Spaces to Support Sense of Place: The Role of Expectation, *CSCW Workshop: The Role of Place in Shaping Virtual Community*, New Orleans, USA, November.

Gaver, W., Smith, R. and O'Shea, T., 1991: Effective Sounds in Complex Systems: The ARKola Simulation, *Proceedings of the CHI Conference on Human Factors on Computing Systems***,** ACM, New York, NY, 85-90.

Gilkey, R. H. and Weisenberger, J. M., 1995: The Sense of Presence for the Suddenly Deafened Adult, *Presence: Teleoperators and Virtual Environments*, **4(4)**, 357-363.

Harrison, J., Rensink, R. A. and van de Panne, M., 2004: Obscuring Length Changes during Animated Motion, *ACM Transaction on Graphics*, **23(3)**, 569-573.

Hughes, C. E., Stapleton, C. B., Micikevicius, P., Hughes, D. E., Malo, S., & O'Connor, M., 2004: Mixed Fantasy: An Integrated System for Delivering MR Experiences. *VR Usability Workshop: Designing and Evaluating VR Systems*, Nottingham, England, January 22-23, 2004 (Proceedings Available on CD).

Larsson, P., Västfjäll, D. and Kleiner, M., 2001: Ecological Acoustics and the Multi-modal Perception of Rooms: Real and Unreal Experiences of Auditory-Visual Virtual Environments, *Proceedings of the 2001 International Conference on Auditory Display, Epoo, Finland*, July 29-August 1, 245-259.

Lombard, M. and Ditton, T., 1997: At the Heart of it All: The Concept of Presence, *Journal of Computer Mediated- Communication* [On-line], **3(2)**. http://www.ascusc.org/jcmc/vol3/issue2/lombard.html

Loomis, J., Klazky, R. and Golledge, R., 1999: Auditory distance perception in real, virtual, mixed environments, In Y. Hota & H. Tamura (Eds.). *Mixed Reality: Merging Real and Virtual Worlds*, Ohmsha, Tokyo, 201-214.

Sekuler, R., Watamaniuk, S. N. J. and Blake, R., 2002: Perception of Visual Motion, In H. Pashler (Series Ed.) & S. Yantis (Vol. Ed.), *Stevens' Handbook of Experimental Psychology: Vol. 1. Sensation and perception (3rd edition).* Wiley&Sons, New York.

Serafin, G. and Serafin, S., 2004: Sound Design to Enhance Presence in Photorealistic Virtual Reality, *Proceedings of the 2004 International Conference on Auditory Display*, Sidney, Australia, July 6-9.

Slater, M., Usoh, M. and Steed, A., 1994: Depth of Presence in Virtual Environments, *Presence: Teleoperators and Virtual Environments*, **3(2)**, 139-144.

Soulodre, G. A., Lavoie, M. C. and Norcross, S. G., 2003: Objective measures of listener envelopment in multichannel surround systems, *Journal of the Audio Engineering Society*, **51(9)**, 826-840.

Turner, P., et al., 2003: Evaluating Soundscapes as a Means of Creating a Sense of Place, *Proceedings of the 2004 International Conference on Auditory Display*, Boston, USA, July 6-9

Wightman, F. L. and Jenison, R. L., 1995: Auditory Spatial Layout, In W. Epstein and S.J. Rogers (Eds.), *Handbook of Perception and Cognition, Volume 5: Perception of Space and Motion*, Academic Press, Orlando, FL.