THE EFFECT OF AUDITORY RENDERING ON PERCEIVED MOVEMENT: LOUDSPEAKER DENSITY AND HRTF

James A. Ballas
Derek Brock
Janet Stroup

Naval Research Lab
Washington, DC, 20375, USA
ballas@itd.nrl.navy.mil

Hesham Fouad

VRSonic, Inc.
Arlington, VA, 22201, USA
hfouad@vrsonic.com

ABSTRACT

Until recently, multiple speaker systems for Virtual Environment (VE) applications were limited to a few front and rear speakers. Utilizing the Virtual Audio Server (VAS) with a Vector Base Amplitude Panning (VBAP) algorithm for multiple speaker control, an array of 24 speakers was constructed to test large speaker configurations. Localization of complex movement was superior with the 24-speaker system, compared to an 8-speaker configuration or HRTF spatialization.

1. INTRODUCTION

Good spatialization of sound is critical for Virtual Environment (VE) applications. While there has been important progress in the development of headphone systems that use perceptual synthesis (i.e, HRTF filtering) for spatialization, speaker systems that simulate a sound field have some advantages [5, 6]. The speaker-based systems have typically utilized 4-5 speakers at a single elevation, and might employ Ambisonics or Dolby 5.1 encoding to simulate the sound field. Recently a 10.2 system has been proposed which adds more front speakers including two upper ones [1]. Speakers are typically used for precise localization research, and one system has an array of 272 speakers mounted within a geodesic sphere [7], with a speaker every 15 degrees horizontally and vertically. Multiple speaker systems are also useful for multiperson, large workspaces as well [2]. Given the advantages of using a speaker to spatialize sounds precisely, an obvious question is whether a large array of matched speakers surrounding a single listener would provide a realistic Virtual Sonic Environments (VSE). We have recently constructed such a system and are in the initial stages of evaluation.

1.1. Overview of System

We are using the Virtual Audio Server (VAS) [3] to investigate issues associated with creating highly realistic VSE. VAS is composed of four independent subsystems: High-level modeling constructs provide a basis for studying new techniques in modeling VSEs. Sound source processing is designed so that new sound representations can easily be integrated into the system. A rendering subsystem enables the seamless integration of novel spatialization techniques without affecting any of the existing design. Finally a scheduling mechanism enables the study of real-time scheduling techniques for the sound generation process.

In this research effort we are primarily interested in evaluating the potential benefits of spatialization using a large speaker array. Thus we extended the rendering subsystem to support control of 24 speakers through three ADAT cards. We had previously implemented the Vector Base Auditory Panning (VBAP) [10] algorithm to control multiple speakers, and only needed to specify the 24-speaker configuration once the ADAT hardware was working. VBAP localizes a sound by choosing an appropriate triplet from the speaker configuration and panning the sound within the triangle formed by the speaker triplet. The clear advantage of VBAP is that it conveys both azimuth as well as elevation information, and it can support an arbitrarily large number of speakers.

1.2. Speaker Configuration and Listening Area

The system used in this study was recently constructed at the Naval Research Laboratory, and to our knowledge, represents the largest speaker array that has been implemented for a single-person VSE and the largest array controlled with the VBAP algorithm. Twenty-four speakers are arranged at three levels around the listener, as illustrated in Figure 1. The distance from the typical head position to the speakers is about 1.37 m. The height at the center of the tall speakers is 2.24 m; the height at the center of the middle speakers is 1.52 m, and the height at the center of the low speakers is 300 cm. Powered, bi-amplified Yamaha MSP5 monitor speakers are used which have a 12 cm two-way cone and a 2.5 cm titanium dome.

The array is placed within a circular enclosure of 16 sound absorption screens (VariScreen™), each 2.44 m high. Each screen was angled to produce an absorption coefficient of about 1.3 for 1-4 kHz. This enclosure provides a subjective dead space with very little perceptible reflections.

Two speaker configurations were used in this first investigation. One of these was the full 24-speaker configuration. The other was a subset of 8 speakers, 4 high and 4 low. In the latter condition, the high and low speakers were offset by 45 degrees in azimuth to accommodate the generation of triangles by VBAP. An 8-speaker configuration with the high and low aligned in azimuth did not function as well as the offset alignment we used. In order to examine how well perceptual synthesis faired in comparison to speaker-based techniques, HRTF spatialization was also used in this investigation. The VAS

system was used to render sounds using HRTF spatialization. The HRTF data sets, however, were not measured for the individual subjects.

The underlying rational for our loudspeaker configurations is based on an analysis of loudspeaker panning in terms of a sampling problem. In loudspeaker panning, loudspeakers represent sample points in the space of all directions to potential sound sources. This space forms a sphere surrounding the listener. Sound source directions that correspond to sample points (loudspeaker locations) can be reproduced exactly, whereas in-between positions are interpolated through panning techniques. By varying the number of speakers used in a panning configuration, the resolution used in sampling the space varies also. Our hypothesis is that increasing the sampling resolution improves the accuracy of the spatialization. In pragmatic terms we wanted to know if increasing the number of loudspeakers in our institution's CAVE facility beyond eight (located at the corners) would be useful.

The speaker configurations used in this experiment approximate, within physical constraints, a uniform sampling of the space of all directions. Other panning approaches such as Dolby 5.1 utilize non-uniform distributions of the sample points to take into account directional variations in spatial acuity of the human auditory system. In many VE applications, however, we cannot make assumptions about the orientation of the listener's head and therefore cannot make such optimizations.

1.3. Movement Perception Paradigm

We are particularly interested in generating moving sounds in VSEs and our initial investigation focuses on this capability. A perceived movement paradigm was used to test the effect of three rendering scenarios on the accuracy of the perceived movement: panning with 8 and 24 speakers, and HRTF spatialization. The stilulus was a synthesized sound which seemed like a fly moving in and out and around the listener. This sound was produced using the following timbre tree:

(combine (* (sinewave 15500) (pulse 100))) (1)

In VAS, this timbre tree produces a 15.5 kHz sine wave modulated by a pulse train of 100 Hz [4]. The free field spectrum measured with a B&K Pulse system, with the microphone set 300 cm from the front of a speaker is shown in Figure 2. The listener's task was to sketch the continuous movement of this object that was perceived while the VAS system modified the speaker output to produce the complex movement pattern shown in Figure 3, which can be thought of as a series of spokes. The height of this movement was kept constant at just above the listener's virtual head position. The pattern was started at different locations for each rendering condition. The pattern took 82 s with minor variation in speed.

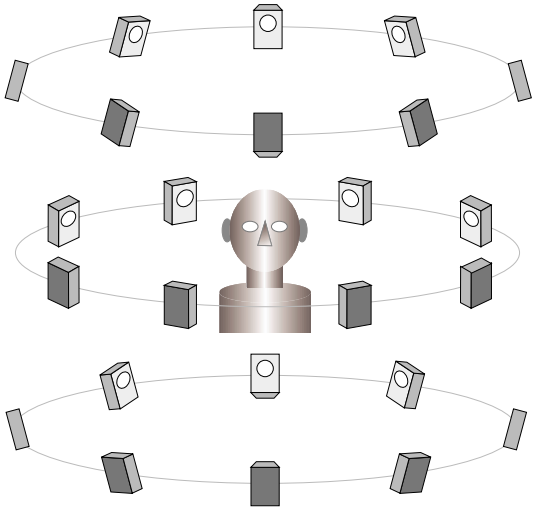

Figure 1. *Twenty-four speaker array.*

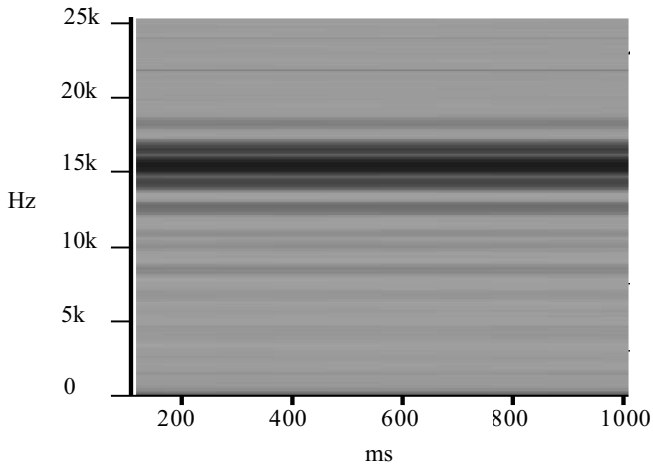

Figure *2. Spectrum of stimulus computed from generated wave*

2. RESULTS

Figures 4 through 7 are scanned images of the patterns drawn by four subjects, listening with the 24 and 8-speaker configurations as well as HRTF spatialization. Our subjective impression is that the patterns produced with the 24-speaker configuration match the controlling spline (Figure 2) better than the 8 speaker patterns or the HRTF patterns.

To test our impression, we overlaid the controlling spline on each of the listeners' sketches and evaluated the degree of match using two types of binary judgements. The first judgment (Area) was whether or not the area of each spoke drawn by the listener fell by more than 50% within the area of the corresponding spoke in the controlling spline or itself covered more than 50% of this spoke. The second judgement (C. Line) was whether the approximate centerline of each spoke drawn by the listener fell within the fan of the corresponding spoke in the controlling spline. Table 1

summarizes our findings. Cumulatively, in the 24-speaker condition there were 36 correct matches (out of a possible 56). There were 31 correct in the 8-speaker condition and 25 correct with HRTF spatialization.

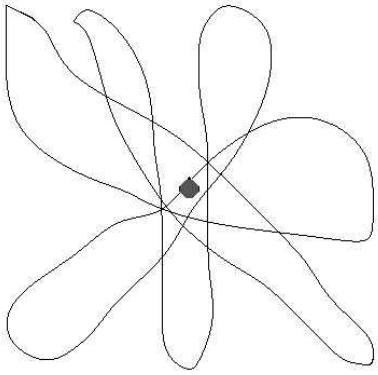

Figure 3. *Spline used to control the movement of the auditory object*

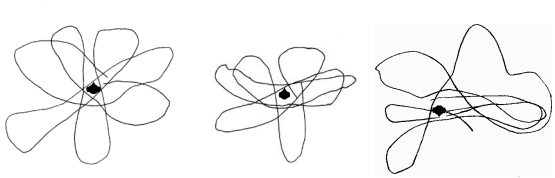

Figure 4. *Movement drawings by Subject 1 with 24-speaker (left), 8-speaker (center) configuration, and HRTF rendering.*

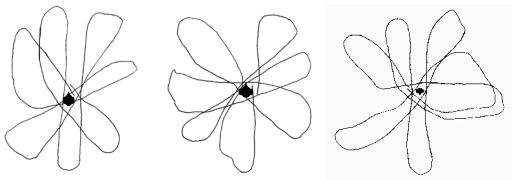

Figure 5. *Movement drawings by Subject 2 with 24-speaker (left), 8-speaker (center) configuration, and HRTF rendering.*

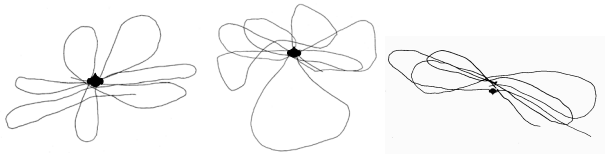

Figure 6. *Movement drawings by Subject 3 with 24-speaker (left), 8-speaker (center) configuration, and HRTF rendering.*

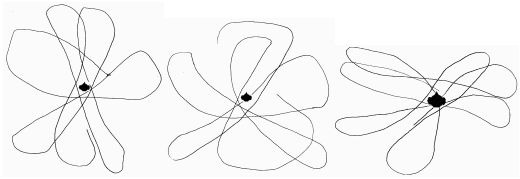

Figure 7. *Movement drawings by Subject41 with 24-speaker (left), 8-speaker (center) configuration, and HRTF rendering.*

Table 1. Accuracy of Perceived Spoke Areas and Centerlines[a].

Spoke	Property	24 speaker				8 speaker				HRTF			
		S1	S2	S3	S4	S1	S2	S3	S4	S1	S2	S3	S4
1:00	Area	1	1	0	1	1	1	1	1	1	1	0	0
	C. Line	1	1	0	1	1	1	1	0	1	1	0	0
3:00	Area	1	1	1	1	1	1	1	1	1	1	1	1
	C. Line	1	1	1	1	1	1	1	1	1	1	1	1
4:30	Area	0	1	0	0	0	1	0	1	1	0	0	1
	C. Line	0	1	0	0	0	0	0	1	0	0	0	1
6:00	Area	1	1	1	0	1	1	1	1	0	1	0	0
	C. Line	1	1	0	0	0	1	1	0	0	1	0	0
7:30	Area	1	1	0	1	1	1	0	1	0	1	0	0
	C. Line	0	1	0	1	0	1	0	1	0	1	0	0
10:30	Area	1	1	0	1	1	0	0	0	0	1	1	0
	C. Line	1	1	0	1	0	0	0	0	0	1	1	0
11:00	Area	1	0	1	0	0	1	0	0	0	1	0	0
	C. Line	1	0	1	0	0	0	0	0	0	1	0	0
Total		11	12	5	8	7	10	6	8	5	12	4	4

[a.] 1 indicates a correct match between the drawing and the controlling spline

To further compare the drawings to the controlling spline, an analysis of the perceived location and direction of each pass by the listener was made. This analysis encoded the location as one of 8 possible orientations, and the direction of the pass as one of eight vector angles. The correctness of the perceptions were tallied and are shown in Table 2, for each of the 7 passes, by subject and by rendering technique. Overall the perceptions were more correct with 24 speakers (45) compared to 8 speakers (34), and were least correct with HRTF (21 correct). A summary table of the results is presented in Table 3. A repeated measures ANOVA produced a significant effect for rendering technique with the summed data from Tables 2 and 3 ($F(2,6) = 6.79$, $p = .03$). The averages for each rendering technique are shown in Figure 8 with standard error bars.

Table 2. Comparison of Perceived Movement Past the Listener[a]

PassBy	Property	24 speaker				8 speaker				HRTF			
		S1	S2	S3	S4	S1	S2	S3	S4	S1	S2	S3	S4
1st	Location	1	1	0	1	0	0	0	0	0	0	1	0
	Direction	1	1	0	1	0	0	0	1	0	0	1	0
2nd	Location	1	1	1	1	1	1	0	1	1	0	0	0
	Direction	1	1	1	1	1	1	0	1	1	0	1	0
3rd	Location	1	1	0	0	1	1	0	0	0	1	0	0
	Direction	1	1	0	1	1	1	0	1	0	1	0	0
4th	Location	1	1	0	0	1	0	0	1	0	1	0	0
	Direction	1	1	0	0	1	1	0	1	0	1	0	0
5th	Location	1	1	1	1	1	1	1	1	1	1	0	0
	Direction	1	1	1	1	1	1	1	0	1	1	1	0
6th	Location	1	1	1	1	1	1	0	1	0	1	0	0
	Direction	1	1	1	1	1	1	0	0	0	1	0	0
7th	Location	1	1	1	0	0	1	1	0	1	1	0	0
	Direction	1	1	1	0	0	1	1	1	1	1	0	1
Total Correct		14	14	8	9	10	11	4	9	6	10	4	1

[a] 1 indicates a correct match between the drawing and the controlling spline

Table 3. Summary of Accuracy of Perceived Movement

	24 speaker				8 speaker				HRTF			
From Table 2	11	12	5	8	7	10	6	8	5	12	4	4
From Table 3	14	14	8	9	10	11	4	9	6	10	4	1
Total by Subj	25	26	13	17	17	21	10	17	11	22	8	5
Total by Technique	81				65				46			

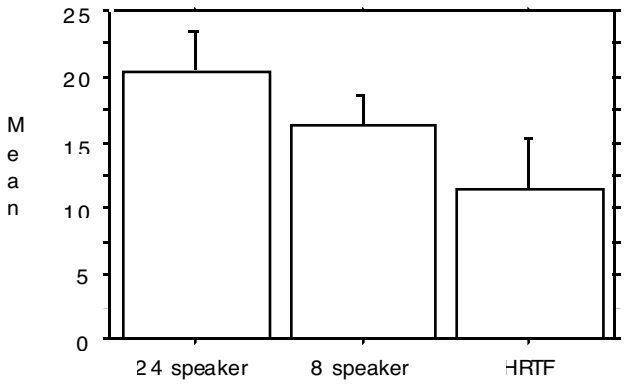

Figure 8. *Mean correct for each rendering technique.*

3. CONCLUSIONS

The results in this first investigation are promising and suggest that a larger configuration of speakers can support more accurate perception of complex auditory motion.

With regard to speaker configuration, one could argue that given *a priori* knowledge of the sound source's path in our experiment (along a lateral plane), a more optimal speaker configuration could have been devised in the 8-speaker case to capture the movement of the sound source. This, however, would have been counter productive since the objective of the experiment was to study the effect of under sampling the space of all directions on the perception of the motion of a sound source.

An examination of the subjects' renderings of the sound source's trajectory in this experiment reveals a surprising outcome. All the subjects perceived the sound source as being rendered, at some point, inside the speaker enclosure near their head. This result is unexpected because according to current thinking on speaker panning [8, 9], the technique is not capable of rendering sounds inside the speaker enclosure. Some experimentation with various sounds revealed that this phenomenon does not occur with all sounds. By listening to a variety of sounds, all following the same trajectory, we observed that only some of the sounds produced this effect. Our initial reaction is to attribute this effect to a psychological process similar to size constancy in visual perception. These observations, however, are very preliminary and further investigation is required to examine this effect.

4. REFERENCES

[1] P. M. Bracke. Breaking the 5.1 Barrier. DVDFILE.com, July 7, 1999, Available at: http://audiolab.usc.edu/DVDFILE_COM.htm.

[2] P. R. Cook, G. Essl, G. Tzanetakis and D. Trueman. N>>2: Multiple-speaker display systems for virtual reality and spatial audio projection. *ICAD'98 conference Proceedings on-line*: http://www.ewic.org.uk., Nov 1-4, 1998.

[3] H. Fouad, J. A. Ballas and D. Brock. An extensible toolkit for creating virtual sonic environments. *Online Proceedings of the International Conference on Auditory Display*, www.icad.org., April 2-5, 2000.

[4] H. Fouad and J. Geigel. The virtual Audio Server Programmer's Guide. Available on-line at ftp.aic.nrl.navy.mil/pub/VAS, 2000.

[5] H. Fouad, (in press) Loudspeaker Panning Techniques, In K. Greenebaum, *Audio Anecdotes*, A K Peters, Natick, MA.

[6] M. J. Evans, A. I. Tew, and J. A. S. Angus. Spatial audio teleconferencing: which way is better? *Proceedings of the International Conference on Auditory Display*, Nov 3-5, 1997.

[7] R. L. McKinley, M. A. Ericson and W. R. D' Angelo. 3-D Auditory displays: Development, applications and performance. *Aviation Space and environmental Medicine*, 65, a31-a38, 1994.

[8] R. Moore, *Elements of Computer Music*, Prentice Hall, Englewood Cliffs, NJ, 1990.

[9] J.P. Pierce, *The Science of Musical Sound*, Scientific American Library, New York, NY, 1983.

[10] V. Pulkki, Virtual Source Positioning Using Vector Base Amplitude Panning. *Journal of the Audio Engineering Soc.*, 1997, vol. 45 no. 6, pages 456-466.