Listening Tests and Evaluation of Simulated Sound Fields Using VibeStudio Designer

György Wersényi

Széchenyi István University,
Hungary
wersenyi@sze.hu

Hesham Fouad

VRSonic,
USA
hfouad@vrsonic.com

ABSTRACT

This paper presents the results of a user-based evaluation of localization accuracy, distance perception as well as room size perception for headphone and loudspeaker based auditory displays. A total of 50 participants listened to four auditory scenes created with VRSonic's VibeStation application. Each scene was rendered using two methods: loudspeaker panning over a 5.0 loudspeaker array and headphone-based spatial sound reproduction using Head Related Transfer Functions (HRTFs). The four scenes were designed to each test a specific aspect of spatial hearing. Scene 1 tested for localization of fixed sources. Scene 2 was used to examine room size perception. Scene 3 was used to test distance perception and Scene 4 tested for localization of moving sources and listener. The participants responded to questions related to the location of each sound they heard as well as transitions between two room sizes and free field. The results of the current study show that the system setup including hardware and software performs as expected and offers a user-friendly way for virtual audio simulation.

1. INTRODUCTION

Virtual auditory displays deal with simulating real world audio experiences [1-3]. Perceiving an auditory event in the real world entails integrating information about both the event itself and its location with respect to the listener. The ability to perceive the spatial location of a virtual sound source entails recreating monaural and binaural cues, and spectral modifications to the acoustic signal reaching a listener [4]. This can be done either through headphone-based spatial sound reproduction using Head Related Transfer Functions (HRTFs) or through multi-loudspeaker panning techniques [5-7].

Head-related Transfer Functions (HRTFs) describe how an auditory event is heard at the human's eardrum [8]. HRTF measurement is an intrusive and time-consuming process and entails playing sounds from designated locations, while recording the sounds using tiny microphones placed inside listener's ears. Individualized recordings of HRTFs are thought to substantially enhance the human's ability to judge sound locations especially when using headphone-based spatial sound reproduction [4]. Due to the complexity of HRTF recording for individual subjects, different catalogues that store HRTF recordings for multiple subjects have been developed; these include AUDIS [9], CIPIC [10], and LISTEN [11].

This paper provides the results of user-based evaluation of sound fields simulated using headphone-based spatial sound reproduction and loudspeaker panning techniques. The objective of the study was to compare subjects' localization accuracy, distance perception, and space perception with sound fields simulated using both these approaches as well as to test the capability of the software environment.

2. BACKGROUND

As aforementioned, spatial audio technology simulates cues that are naturally present and enable listeners to locate sounds in the real world. More specifically, humans perceive sound location in three dimensions; azimuth, elevation, and distance.

Interaural time and intensity differences (ITD and IID) are used for localizing a sound source's angular position (azimuth). Interaural cues are based on the relative difference between wave fronts at the two ears on the horizontal plane [5, 12]. IID and ITD, do not however provide sufficient information for a listener to disambiguate between source positions in the frontal hemisphere and corresponding positions in the rear hemisphere. This is because IID and ITD values are identical for a given position in one hemisphere and its reflected position in the other ("cone of confusion").

The human pinnae provide spectral modifications to the acoustic signals that aid in both disambiguating front and back sources as well as elevation judgment with respect to the median plane [13]. The spectral modifications resulting from pinnae folds produce a unique set of micro-time delays, resonances, and

diffractions that translate into a unique descriptor for each sound source position in the median plane [4]. These spectral modifications are particularly important for modeling the HRTF of a listener.

The intensity of a sound source is the most prominent distance cue in anechoic environments (or with familiar sounds) [14, 15]. The intensity of a sound is inversely proportional to the squared distance from the sound source. In reverberant environments the ratio of reflected to direct sound plays an important role for distance perception [5], this ratio creates perceptual differences in the sound quality that depend on source distance [15].

HRTF-based spatial audio reproduction deals with modeling the acoustic signal modifications resulting from a listener's head, torso, and pinnea reflections. HRTF measurements entail placing tiny microphones inside the listener's ear canal. Then sounds are played from an array of loudspeakers precisely placed at known locations around the listener [16]. When the sounds are played, examining the spectral difference between the known played sound and the sound recorded by the microphones enables the extraction of the modifications that are unique to the listener. These spectral modifications are then stored and can be used to play sounds to a listener. It is important to note that the proper choice of HRTF is crucial to truly simulate sound source positions. For example using a non-good sound localizer HRTF can worsen that of a naturally good sound localizer [17].

HRTF-based sound reproduction is best if individualized HRTFs are used. In one study it was found that localization accuracy using headphones (and individualized HRTFs) resulted in comparable performance to free field listening (i.e., localization blur of about 5-10 degrees), nevertheless the rate of front-back confusions increased from 6% to 11% and elevation judgments became less defined [18]. Using non-individualized HRTFs, ITD and IIDs are synthesized but some spectral information is distorted, which leads to ambiguous elevation judgments and increased front-back errors [19].

The other approach used for sound field simulation is the use of free field loudspeaker arrays with either amplitude panning or wave field synthesis approaches. Loudspeakers strategically placed around a listener can be used to simulate the angular location of a sound source by manipulating the signals being played over loudspeakers. Panning approaches simply scale the amplitude of a sound signal presented over two (2D arrays) or three (3D arrays) loudspeakers to give the impression of a positional source. Most surround sound implementations utilize this approach. The other approach, wave field synthesis, attempts to recreate the incident wave front of a source at the listener using a large number of loudspeakers arranged in a line-array configuration. This approach, while producing good results, requires a very large number of loudspeakers to be effective.

3. METHOD

3.1. Participants

A total of 50 subjects participated in the listening tests, 13 females and 37 males. The minimum age was 18; the maximum age was 50 with a mean value of 29.5 years. Table 1 shows how frequently subjects use headphones.

Daily	Several times a week	Several times a month	Seldom, never	
9	15	18	8	Number of subjects

Table 1. User headphone usage routine.

3.2. Apparatus

The experimental setup consisted of a usual desktop computer equipped with a Creative Audigy sound card and an external TerraTec Aueron 5.1 MK II USB sound card providing 6-channel analog outputs. The loudspeaker display consisted of 5 loudspeakers positioned in a typical surround sound configuration: front-left, center, front-right, surround left and surround right similarly to Fig.2. The JMLAB CC700 was used for center speaker and four Chorus 707 speakers for the rest. All five are 2-way bass-reflex systems with a frequency response of about 60 Hz to 22 kHz. The analog outputs of the TerraTec sound card were connected to the external inputs of a DENON AVR-3805 home theater receiver. The listening room was a nearly empty, large rectangle room with an average reverberation time of 0,8 sec. Subjects were instructed to keep their head still during the listening tests.

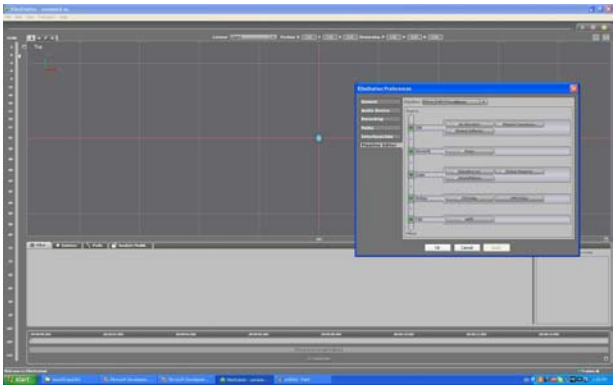

Figure 1. VibeStation application with audio pipeline editor displayed.

Headphone playback was done over a pair of AudioTechnica ATH-D40fs circumaural headphones connected directly to the computer's audio card.

The simulated sound fields were created using VRSonic's VibeStudio Designer software suite [20]. VibeStudio Designer consists of the VibeStation application for spatial audio scene design and the Profiler application for HRTF selection based on a best-fit selection method [12]. VibeStation is capable of rendering scenes over 2, 4, 5 and 7 loudspeakers and over headphones using binaural synthesis with HRTFs. Larger loudspeaker arrays (up to 48 loudspeakers) can also be supported with the addition of a SoundSim Rack external rendering appliance that interfaces with the VibeStudio applicaton. The software allows users to configure the audio rendering pipeline by including and excluding processing stages in the audio pipeline and by selecting rendering algorithms for loudspeaker panning (Fig. 1).

The Profiler application guides the user through a selection process that results in a stored listener profile. Listener profiles specify the user's interaural distance, head tracker offsets and HRTF dataset selection. By default the program provides 7 HRTF datasets from the CIPIC and LISTEN catalogues. These include CIPIC subjects 3, 8, 9, 10, and 11; LISTEN subject 3 and a generalized HRTF dataset. The full CIPIC and LISTEN catalogues can be downloaded resulting in 97 HRTF datasets that can be selected.

Rendered scenes can be recorded for later playback and editing as either a single, multichannel audio file or multiple, single channel audio files. The stereo single file format is well suited for playback over headphones or stereo loudspeaker setups without running the software. Multichannel playback, however, can be realized only while running the software with the appropriate loudspeaker setup.

3.3. Experimental Design

For the listening tests we created four scenes using the VibeStation application. The scenes were rendered over both headphones and loudspeakers at approximately the same loudness. Each participant was presented with both playback methods and the results were compared. For the 5.0 loudspeaker display we selected the Vector Based Amplitude Panning (VBAP) loudspeaker-panning algorithm. For the headphone display we selected the CIPIC "subject 3" HRTF dataset.

Scene 1 used the sound of a ringing telephone. Source locations were positioned 45 degrees around the virtual listener's head (Fig. 2). The playback order was randomized in 6 seconds intervals. The task was to identify the source locations.

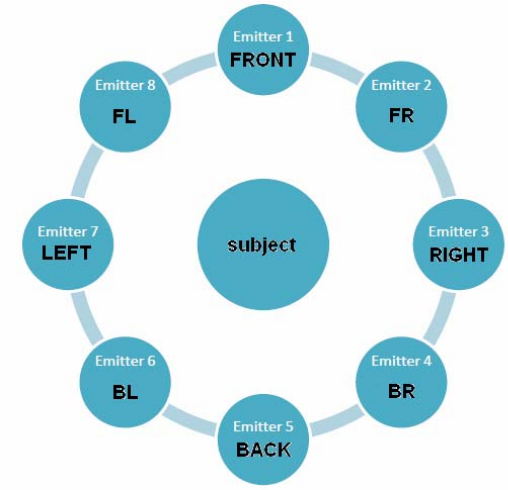

Figure 2. Sound source locations for scene 1. FL, FRONT, FR, BL and BR are also actual loudspeaker positions.

Scene 2 used looped music as the virtual listener moves in the sound field from the free field into a smaller room, then again into the free field and finally into a larger room. The task was to detect the transitions and to estimate room size (which one is small and big). The smaller room was set to 15 x 4.5 x 2 meters whilst the bigger one was set to 20 x 20 x 10 meters, but all other parameters were the same (perfect reflectors material).

Scene 3 used the sound of a honk of a car in front of the listener. The distance first was simulated 40 meters (100%) then it was decreased to 20 meters (50%) and again to 10 meters (25%). The task was to detect that the distance was decreased to the half every time. Finally, we asked the subjects to make a raw estimate in meters.

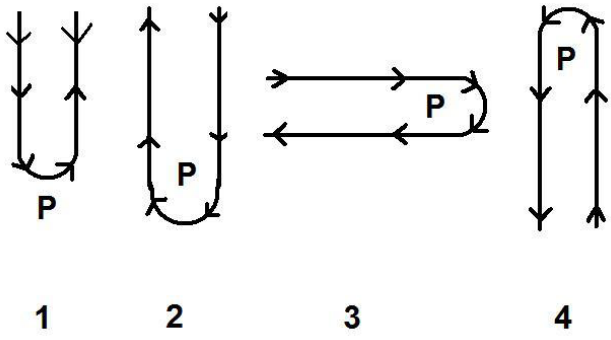

Figure 3. Set of possible trajectories. P indicates the listener's position.

Scene 4 included a trajectory of a flying object. For 5.0 loudspeaker playback we used the sound of a helicopter, for headphone playback we used the sound

of an airplane. The task was to select the proper trajectory from a set of four different possibilities as shown in Fig 3.

3.4. Experiment Procedure

Prior to the start of an evaluation session, each participant completed an informed consent and a demographics questionnaire. A detailed explanation of the measurement process was given. Each subject listened first to scene 1 using randomized presentation order of sound sources. This was followed by scenes 2 to 4. After each scene questions were answered referring to that scene. The measurement was about 30 minutes. Measurements with the loudspeaker setup were executed in the university laboratory at a later time. The same 50 subjects participated in this test.

3.5. Evaluation

3.5.1. Headphone Playback

Scene 1

Table 2 summarizes the results of subjects' localization accuracy with the headphone rendering of Scene 1. The diagonal indicates correct answers. There are no left-right reversals but front-back reversals are frequent. Front-back reversals are one of the main problems in virtual and sometimes in real life localization [21, 22] This is also present on the sides where, for example, front-left is confused with back-left. Subjects often described back sources as frontal sources with lower loudness level.

%	Front	FR	Right	RB	Back	BL	Left	FL
Front	**80**	8			43			4
FR	12	**52**	21	20	2			
Right		34	**69**	19				
RB		6	10	**61**	2			
Back	4				**53**			
BL						**61**	14	18
Left						16	**66**	24
FL	4					23	20	**54**

Table 2. Results of Scene 1 with headphone playback. Compare with Table 3.

Scene 2

The recognition of spatial properties was nearly perfect, only 3 subjects failed. Both the transitions as well the room size estimation were easy tasks for the subjects. Only 6 people thought that the first room would be bigger. These decisions were based on the simulated reverberations. Because both rooms were highly reflective environments (metal-like), the differences between transitions were easy to detect. Setting different room sizes or materials to create smaller differences in reflections could result in larger errors.

Scene 3

The first drop (from 40 to 20 meters) in the distance was detected correctly by 75% of the participants, while the second drop (from 20 to 10 meters) was detected correctly by only 62% of the participants. We expected that the estimation of the distance in meters would result in a wide range of numbers. The task was to estimate the middle source position that is simulated at a distance of 20 meters. About 30% could give a relatively good estimation of the distance, 50% estimated the distance as being further than it was (50-100 m) and 20% estimated the distance to be closer than 5 meters. This result is expected as distance perception depends on a variety of cues including familiarity with a sound, the ratio of direct to reverberant energy reaching the listener, and spectral changes to the source. In this scenario the only cue present for detecting distance was spreading loss.

Scene 4

The best performing simulated trajectory was number 2 (Fig. 3), 82% detected it correctly. Subjects were allowed to listen to the sound three times. The mean value for the number of auditions was however only two. We observed that people who seldom or never use headphones needed three auditions. In case of incorrect localization, subjects usually guessed trajectory 3.

In general, younger people (20-27 years of age) and frequent headphone users were better almost in every task. Only in front-back confusions are results independent from gender, age or headphone user routine.

For test with personalized HRTF we had only 10 subjects. Personalization means setting the head diameter and physical properties for a better interaural time difference simulation using the Profiler application. The HRTF used in these conditions was the same as before (subject 3 of the CIPIC database). Seven subjects had the same results with and without personalization. One had worse and two had better results with personalization (decreased rate of front-back confusion). These results are only informal due to the small number of participants.

3.5.2. Loudspeaker playback

Scene 1

%	Front	FR	Right	RB	Back	BL	Left	FL
Front	**86**	8						14
FR	9	**78**	25					
Right		14	**66**	14				
RB			9	**74**	20			
Back				12	**66**	15		
BL					14	**76**	21	
Left						9	**65**	11
FL	5						14	**75**

Table 3. Results of Scene 1 with loudspeaker playback. Compare with Table 2.

Results were overall better for loudspeaker playback compared to headphone listening. It was very helpful that the physical positions of the loudspeakers were identical to the simulated virtual directions in five cases, and sound was only transmitted from the actual loudspeaker (Fig. 2.). In the three cases where the virtual source did not coincide with a loudspeaker position, the virtual sound source was created by two loudspeakers (LEFT, RIGHT, BACK). Front-back confusion disappeared, a symmetrical diagonal can be seen in Table 3. Correct judgments are around 65-86%. In case of localization errors subjects mentioned one of the closest virtual positions. This fact is reflected by the diagonal of Table 3. (e.g if the simulated source was FR, incorrect answers included only FRONT and/or RIGHT).

Scene 2

Surprisingly, in Scene 2 the results for loudspeaker playback were the same as headphone playback: only 3 subjects failed to detect the transitions, and only 5 subjects failed to detect the correct room size. We have to take into account that the listening room play a significant role and different listening rooms could result is different results.

Scene 3

The first drop (from 40 to 20 meters) in the distance was detected correctly by 76%, the second (from 20 to 10 meters) only by 65%. About 18% could give a relatively good estimation. 54% of the rest estimated it too far (50-100 m) and 28% estimated it closer than 5 meters. This is almost the same as by headphone playback.

Scene 4

Subjects performed best with simulated trajectory number 3 (Fig.2.). 74% of the subjects detected it correctly, this is slightly worse than the headphone playback condition. It was helpful that sounds from behind come actually from real loudspeakers behind the listener. Subjects could listen to the sound three times and the mean value for the number of auditions was again two.

In general, younger people (20-25 years of age) are better almost in every task. It is important, how the loudspeakers are positioned and what kind of virtual source directions will be simulated. Front-back confusion is not present, mainly due to the center loudspeaker. It seems to be a good idea to use a center speaker. Listeners suggested that room size detection was easier via headphones, maybe due to the listening room properties during loudspeaker playback.

4. DISCUSSION AND CONCLUSIONS

50 subjects participated in a listening test using headphone playback and loudspeaker setup. Headphone playback included non-individual HRTF synthesis while loudspeaker setup used a 5.0 installation. For both tests four different scenes were rendered to test localization, front-back reversals, distance estimation and room models using VRSonic's VibeStudio Designer. The software environment allows easy access to parameters and controlling the simulation. Results of the listening tests are comparable to former results in the literature.

5. FUTURE WORK

Some considerations about the program and future planning:

- There is no built-in wave editor in VibeStation. Using VibeStation and a wave editor in parallel can sometimes be blocked by the ASIO driver. Other drivers may work parallel.
- The "emitter database" is very small, there are only two built-in wave files. This means, one has to download, record and edit the wave files.
- Adding measured, individual HRTFs to the HRTF database requires that the measured HRTFs be converted into the program's SAF format. There were no tools provided with the program to do this.
- Rooms are very simple, geometrical forms, there is no CAD option and it is a simple reverberation simulation for the room only.

- The distance model could be extended by some low-pass filtering that simulates air absorption. This function is implemented in the current version of the software.

6. ACKNOWLEDGEMENTS

This paper is based upon work supported in part by the Office of Naval Research (ONR). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views or the endorsement of ONR. We also thank the participants involved and Christian Herrmann at the University of Applied Sciences, Leipzig, Germany for programming and leading the listening tests.

7. REFERENCES

[1] G. Kramer, "An introduction to auditory display". In G. Kramer (Ed.), *Auditory Display* (pp. 1-77). Reading, MA: Addison-Wesley, 1994.

[2] B. Gygi, V. Shafiro, "From signal to substance and back: insights from environmental sound research to auditory display design". Proc. of ICAD'09. pp. 240-251, 2009.

[3] M. Kleiner, B. I. Dalenbäck, P. Svensson, "Auralization – an overview." *J. Audio Eng. Soc*. v*ol. 41*, pp. 861-875, 1993.

[4] D.R. Begault, *3D Sound for Virtual Reality and Multimedia*. Academic Press, Inc., Cambridge, MA, 1994.

[5] J. Blauret, *Spatial hearing: The psychophysics of human sound localization*. Translated by J.S. Allen. MIT Press, Cambridge, Mass, 1983.

[6] W. Ahnert, S. Feistel, T. Lentz, C. Moldrzyk, S. Weinzierl, "Head-Tracked Auralization of Acoustical Simulation". Preprint 6275, 117th AES Convention San Francisco, 2004.

[7] S. Ferguson, D. Cabrera, "Vertical Localization of Sound from Multiway Loudspeakers". *Journal of the AES, vol. 53(3)*, pp. 163-173, 2005.

[8] E.M. Wenzel, M. Arruda, D.J. Kistler, & F.L. Wightman, "Localization using Non-individualized Head-related Transfer Functions", *Journal of the Acoustical Society of America, vol. 94*, 1993, pp. 111-123.

[9] https://www.european-acoustics.org/Products/ Documenta/Publications/09-de2

[10] http://interface.cipic.ucdavis.edu/CIL_html/CIL_HRTF_database.htm

[11] http://recherche.ircam.fr/equipes/salles/listen/

[12] B.U. Seeber, & H, Fastl, Subjective Selection of Non-Individual Head-Related Transfer Function, Proc. 2003 *International Conference on Auditory Display*, pp. 259-262, Boston University, Boston, MA, July 6-9, 2003.

[13] J. Hebrank, & D. Wright, "Spectral cues used in the localization of sound sources on the median plane", *Journal of the Acoustical Society of America ,vol.* 56, pp. 1829-1834, 1974.

[14] P. McGregor, A.G. HORN, & M.A. Todd, "Are familiar Sounds ranged more accurately?" *Perceptual and Motor Skills,* vol. 61, 1082, 1985.

[15] J.C. Middlebrooks, & D.M. Green, " Sound localization by human listeners", *Annual Review of Psychology, vol.* 42, pp. 135 – 159, 1991.

[16] F.L. Wightman, & D.J. Kistler, "Headphone simulation of free-field listening I: Stimulus synthesis", *Journal of the Acoustical Society of America, vol.* 85, pp. 858-867, 1989.

[17] E.M. Wenzel, "Localization in Virtual Acoustic Displays, *Presence, vol.* 1, pp. 80-107, 1992.

[18] F.L. Wightman, & D.J. Kistler, "Headphone simulation of free-field listening I: Psychophysical validation", *Journal of the Acoustical Society of America, vol.* 85, pp. 868-878, 1989.

[19] E.M. Wenzel, M. Arruda, D.J. Kistler, & F.L. Whitman WIGHTMAN, "Localization using Non-individualized Head-related Transfer Functions", *Journal of the Acoustical Society of America, vol.* 94, pp. 111-123, 1993.

[20] http://www.vrsonic.com/products/vibestudiodesigner.html

[21] D. R. Begault, E. Wenzel, M. Anderson, "Direct Comparison of the Impact of Head Tracking Reverberation, and Individualized Head-Related Transfer Functions on the Spatial Perception of a Virtual Speech Source". *J. Audio Eng. Soc. 49(10),* pp. 904-917, 2001.

[22] P. A. Hill, P. A. Nelson, O. Kirkeby, "Resolution of front-back confusion in virtual acoustic imaging systems". *J. Acoust. Soc. Am. 108(6)*, pp. 2901-2910, 2000.