Audio Engineering Society

Convention Express Paper 173

Presented at the 155th Convention

2023 October 25-27, New York, USA

This Express Paper was selected on the basis of a submitted synopsis that has been peer-reviewed by at least two qualified anonymous reviewers. The complete manuscript was not peer reviewed. This Express Paper has been reproduced from the author's advance manuscript without editing, corrections or consideration by the Review Board. The AES takes no responsibility for the contents. This paper is available in the AES E-Library (http://www.aes.org/e-lib) all rights reserved. Reproduction of this paper, or any portion thereof, is not permitted without direct permission from the Journal of the Audio Engineering Society.

Binaural renderers accuracy comparison: Part I

Lisa LaFountaine, Raymond Plasse, and Dr. Wesley Bulla

Belmont University, 1900 Belmont Blvd, Nashville, TN 37212

Correspondence should be addressed to Lisa LaFountaine (`Lisa.LaFo@aol.com`) and Raymond Plasse (`rayplasse@comcast.net`)

ABSTRACT

This two-part study explored the efficacy of binaural renderers to accurately reproduce the placement of objects within a three-dimensional, virtual soundscape. Many previous works have only tested localization on the horizontal plane (Part I) whereas this research expanded on prior methodology by adding vertical targets along the medial and two sagittal planes (Part II). Two industry leading binaural renderers were compared. The subject task was to map where each sound source was perceived onto a planar response sheet. Results were consistent with previous research in that renderer performance was found to be weak in the horizontal domain. Findings presented here support the notion that horizontal plane localization cannot be solely relied upon to assess the quality of binaural renderers. In part two, further analysis of loci along the medial and sagittal planes will provide a more complete understanding of renderer performance and areas for potential improvement.

1 Introduction

Immersive audio has played a crucial role in three-dimensional sound experiences as applied to film and virtual- or augmented-reality (VR/AR) gaming. Today, immersive audio deliverables have become music, film, broadcast, and streaming standards, providing new opportunities for creators and artists to express their creativity. Optimal playback of these deliverables involve loudspeakers encircling the listener in a so-called “surround” arrangement with added channels above or even below the listener. Examples of such formats include: Dolby Atmos 7.1.4, Auro 3D [1], Sony 360º [2], and NHK 22.2 [3]. While public venues equipped with these immersive formats are growing in numbers, for personal and home use consumer access is often costly. To recreate such spaces at home is similarly neither practical nor affordable for most individuals [2].

Headphones offer an economical solution without the need for specialized hardware, but whose quality depends heavily on convincing signal processing to render spatial audio in a way that sounds natural to listeners [4]. To succeed as a viable playback medium, headphones must accurately reproduce spatial cues found in immersive audio content across two channels without losing any important information. This process, known as “binaural rendering”, essentially simulates an acoustic environment by encoding psychoacoustic properties of head related transfer functions (HRTFs) via digital signal processing [5]. For immersive mixes to translate well through headphones, binaural rendering algorithms utilize generalized HRTF filters to model how sound would arrive at a listener's ears within a three-dimensional acoustic space [6]. An HRTF signature contains temporal, dynamic, and spectral modifications introduced by the ears, head, and torso. Such spatial information reveals essential localization and spatial cues to the listener which gets encoded into the binaural rendering algorithm. This HRTF processing takes into account interaural time and level differences (ITD and ILD) for sound sources outside the medial plane and spectral shape cues for sources along the medial plane [7]. Due to physical differences between ear shapes, HRTFs vary considerably from listener to listener thus degrading the effectiveness of binaural renderers utilizing generalized HRTF signatures [8].

In music, film, and VR/AR, immersive mixes rendered for binaural playback are intended to generate a sense of “envelopment”, “realism”, and “presence.” Consequently, most research has concentrated on evaluating qualitative aspects of immersive formats, appealing to subjects’ emotions and preferences [5, 9 –12]. In many of these reports, binaurally rendered material has not consistently garnered the favor of its subjects. This perceptual trend is reflected among consumers, who often prefer traditional stereophonic playback due to timbral distortion and spatial width problems caused by standard HRTF processing [12].

Reardon et al. [13] tested six different binaural renderers and analysed how accurately each performed the task of eliciting static localization cues along the horizontal plane. Results revealed content-dependent inconsistencies across renderers with cue perception afflicted by localization blur and front/back confusions. The authors concluded that a comprehensive evaluation of the performance of binaural renderers would ultimately inform their subjective appraisal for immersive audio content. While the presented results give some promise that binaural rendering can elicit a localization percept on all sides of a listener, subject responses were limited to sources positioned along the horizontal plane.

The present study, modelled after [13], focused on evaluating localization differences between two industry leading binaural renderers, i.e., their efficacy to accurately reproduce virtual sound source location over headphones. This paper presents an exclusive analysis of renderer performance in the horizontal domain with analysis of the medial and sagittal plane data to follow in part two. As more audio professionals transition to immersive mixing, the ability of binaural renderers to effectively recreate immersive content remains an important topic of investigation.

2 Methods

This study investigated static binaural localization along the horizontal plane, the medial plane, and two sagittal planes. The authors define the medial plane as starting directly in front of the listener, proceeding overhead, and ending directly behind the listener. The left and right sagittal planes take a similar path, diagonally intersecting the horizontal and medial planes as shown in Fig. 1. Stimuli consisted of three consecutive 1500ms pink noise bursts separated by 500ms of silence. For testing purposes, using the most robust signal possible was an integral component of the study since spectral variation introduces mixed results when determining localization accuracy [7].

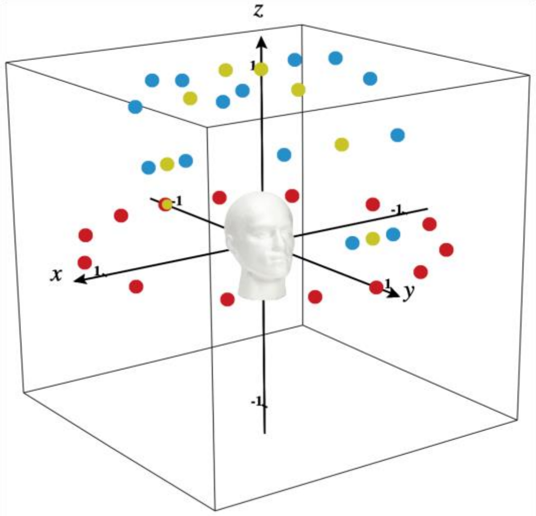

Fig. 1: Spatialized stimuli along the horizontal, medial, and left and right sagittal planes.

Using the “spatial panner” (Fig. 2) accompanying each renderer within a digital audio workstation, each series of bursts was assigned unique azimuth and elevation coordinates along a virtual hemisphere surrounding the listener. Thirty-six target positions were created at or above the listener’s ear level: fourteen along the horizontal plane, eight along the medial plane, and seven positions each along the left and right sagittal planes (Table 1). Each target resided in the center of a zone ranging from 20° to 30° in width; a greater range of degree values was afforded to the side regions to accommodate for a larger minimum audible angle (MAA) as defined in [14]. Instead of accommodating for a larger MAA in the overhead regions, the opposite process was applied to medial targets consisting of more concentrated zones directly above the listener, which was reflected in the corresponding left and right sagittal targets. All thirty-six positions were panned, routed through each renderer (REN-A and REN-B), and printed for a total of seventy-two individual stimuli. These stimuli were randomized and presented across two sets of trials for a total of 144 data points collected per subject.

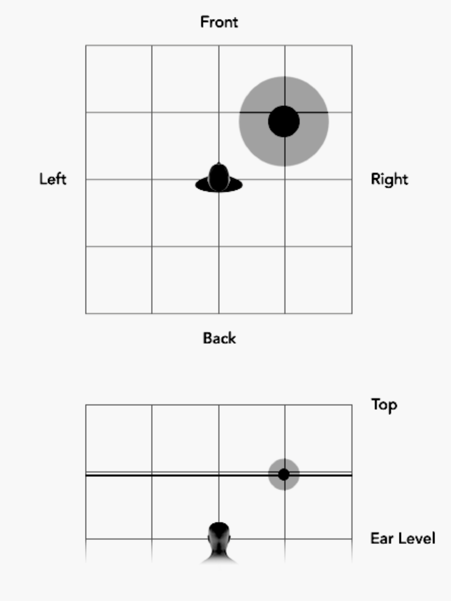

Fig. 2: "Spatial panner" utilized for assigning azimuth and elevation coordinates to stimuli.

Each subject was asked to map the perceived location of the three-dimensional sound source using a hash mark onto two circles, one representing the horizontal plane, or degrees of azimuth, and the other representing the medial plane, or degrees of elevation. In cases where a source was only perceived on one plane, either horizontal or medial, subjects were directed to cross out the unused plane with an "X" (Fig. 3). Thus, a sound source perceived along either sagittal plane would have two corresponding hash marks representing both the appropriate azimuth and elevation for that source (Fig. 4).

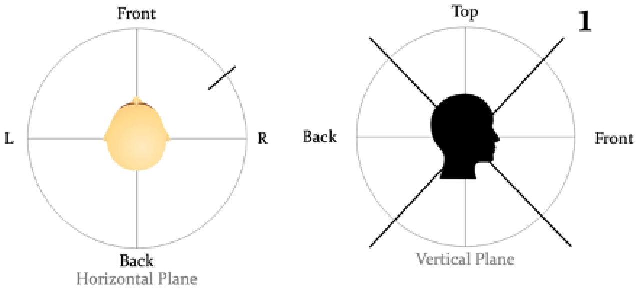

Fig. 3: Subject response sheet showing localization percept for a horizontal plane target (45º, 0º).

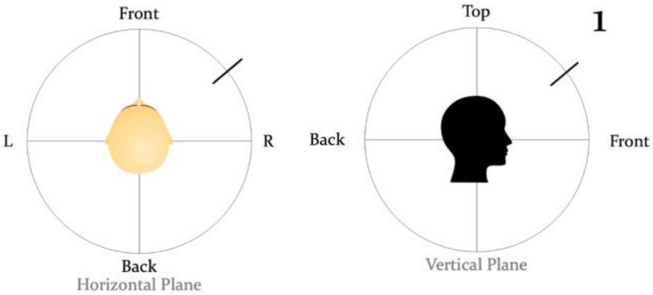

Fig. 4: Subject response sheet showing localization percept for a right sagittal plane target (45º, 45º).

Table 1: Selected zones and tested position in degrees for horizontal, medial, and left and right sagittal planes.

Plane	Zone & Range	Target Position
Horizontal	1 (350° - 10°)	0°
	2 (10° - 30°)	20°
	3 (30° - 60°)	45°
	4 (60° - 90°)	75°
	5 (90° - 120°)	105°
	6 (120° - 150°)	135°
	7 (150° - 170°)	160°
	8 (170° - 190°)	180°
	9 (190° - 210°)	200°
	10 (210° - 240°)	225°
	11 (240° - 270°)	255°
	12 (270° - 300°)	285°
	13 (300° - 330°)	315°
	14 (330° - 350°)	340°
Medial	1 (0° - 30°)	15°
	2 (30° - 60°)	45°
	3 (60° - 80°)	70°
	4 (80° - 100°)	90°
	5 (100° - 120°)	110°
	6 (120° - 150°)	135°
	7 (150° - 170°)	165°
	8 (170° - 190°)	180°
Right/Left Sagittal	1 (0° - 30°)	15°
	2 (30° - 60°)	45°
	3 (60° - 80°)	70°
	4 (80° - 100°)	90°
	5 (100° - 120°)	110°
	6 (120° - 150°)	135°
	7 (150° - 180°)	165°

Fourteen college students ranging from 20 – 30 years of age participated in the listening experiment. Subjects here are considered more experienced than "novice" but not yet "expert" listeners. Each participant reported normal hearing, completed a graduate-level course in critical listening, and had some level of audio engineering and production experience. Subjects were unaware of the purpose of the experiment and received no prior training. Each trial was administered in a sound-treated classroom using Sennheiser HD 595 open-back, circumaural headphones. The experiment, including rest periods, took approximately ninety minutes to complete.

While the full scope of the experimental design is presented above, the following section provides an analysis of horizontal plane responses alone.

3 Results

Data points were resolved to the degree for each response. Front/back confusions were identified and removed for separate analysis. The data were scored two ways: (1) mean number of successes within each target zone and (2) mean absolute error in degrees from each target position. Analyses indicated REN-A data were not gaussian ($\chi2 = 11.6$, $p < .001$, $N = 72$). Therefore, non-parametric (Freedman and Wilcoxon) analyses were used.

Matched pairs tests found no significant difference between REN-A ($M = 0.33$, $SD = 0.20$) and REN-B ($M = 0.31$, $SD = 0.12$) horizontal plane scores ($Z = 0.91$, $N = 14$, $p = .363$), indicating that neither REN-A nor REN-B performed considerably better than the other. For degrees of error, similar trends were observed for influence of presentation plane, again revealing no significant difference between REN-A ($M = 33.7$, $SD = 17.4$) and REN-B ($M = 37.6$, $SD = 14.0$) horizontal plane scores ($Z = 0.97$, $N = 14$, $p = .331$). This result confirms both renderers contributed similar error in subject responses from horizontal targets.

ANOVA revealed significant main effect for horizontal front/back confusions occurring between REN-A and REN-B ($\chi2$ (3, $N = 14$) = 31.4, $p < .001$) with more confusions reported in the back than the front. No significant differences were found between REN-A ($M = 0.37$, $SD = 0.27$) and REN-B ($M = 0.36$, $SD = 0.24$), but rather, a systematic difference was observed between front/back reversal rates as a whole. Stimuli presented at the back of the head were more likely to be perceived in the front whereas the converse was not necessarily true.

4 Discussion

Results were consistent with previous research in that correct judgments for horizontal targets were relatively low and front/back confusions appeared to be prevalent. One might expect localization cues in the horizontal domain to be salient considering interaural differences (ITD and ILD) are generally consistent across listeners [7]. Nevertheless, subjects struggled to consistently map ear-level percepts, suggesting horizontal location errors in reproduction stems from a fundamental weakness in binaural renderer processing.

Fig. 5 depicts the average number of correct judgments per target zone indicating listeners could best localize sound sources from directly in front as well as directly to the right and left of the head. Fig. 6 shows the average absolute error of subject responses from the target indicating between 20° – 40° of average angle error for almost all zones apart from zone 14.

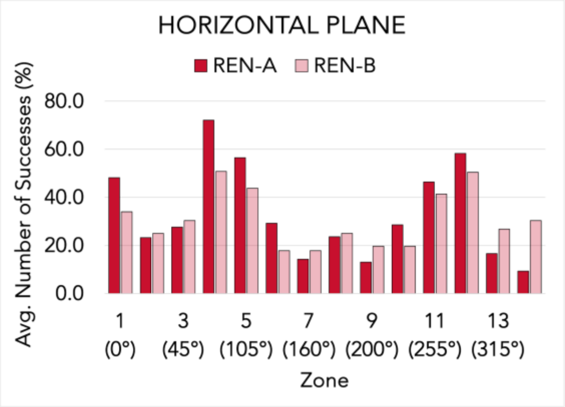

Fig. 5: Average number of successes for each renderer along the horizontal plane corresponding with zones depicted in Table 1.

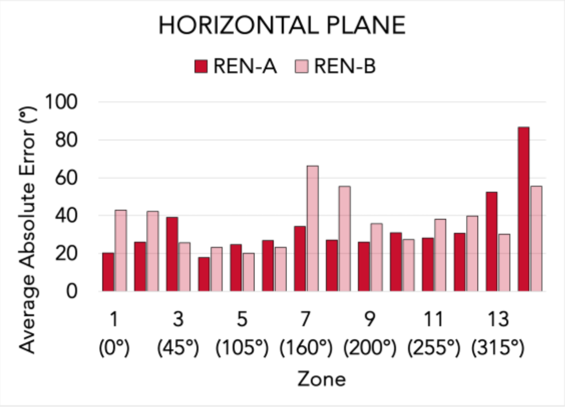

Fig. 6. Average absolute error in degrees for each renderer along the horizontal plane corresponding with zones depicted in Table 1.

Prior research has commonly included training as a way for participants to become more comfortable completing the task of sound source localization with efficiency [6]. Training listeners is often appropriate when specific anomalies or artifacts of a target signal are known and identifiable. “Front-back reversals” (FBR) and “Inside-the-head” reproductions (ITHR) are the two primary types of identifiable distortions that commonly occur with static binaural reproduction [15]. FBR and ITHR were perceptual results-of-interest of the study design and need no such training as they are easily identifiable from specific patterns of misidentified locus origins. FBR and ITHR resolution, or the lack thereof, were part of the error in which the experiment was designed to reveal. In the case of ITHR perception, training could have contaminated data with erroneous responses that do not represent the true perceptual response of the subject. For example, ITHR percepts are often reported as elevated signals that appear to rise from the region of the ear to above (or into the center) of the head. As such, a listener could be trained to report an elevated above-the-head percept as the front-center stimulus thus generating useless data in both the horizontal and elevated domains.

The authors of [13] claim that “assessing localization at elevations other than zero would prove to be difficult and possibly an unfair measure of binaural rendering quality.” However, excluding elevated loci raises some suspicion regarding the possible effect of certain identifiable distortions on the data intended for analysis. Forfeiting those data from analysis yields an incomplete picture of renderer reproduction performance. Future analysis of medial and sagittal plane data will support examination of identifiable distortion phenomena and their role in determining binaural renderer localization accuracy.

5 Conclusion

The outcomes of this experiment were consistent with findings presented in [13] for evaluation of the reproduction accuracy of binaural rendering along the horizontal plane. Considering those results were reported over five years ago, the current state of binaural renderers has shown little improvement and leaves more to be desired. Analysis of medial and sagittal plane data in part two of this study will allow for a clearer picture of the current state of binaural rendering to be presented.

6 Acknowledgements

We would personally like to thank Dr. Wesley Bulla, our third author, for his expert data analysis, master editing skills, and general can-do attitude; Dr. Song Hui Chon and Dr. Doyuen Ko for inspiring this research; Belmont University, the Mike Curb College of Entertainment & Music Business, and our fellow graduate students for allowing us to utilize their facilities, equipment, and ears; and our anonymous peer reviewers for their insight and encouragement.

References

[1] R. J. Ellis-Geiger, "Music Production for Dolby Atmos and Auro 3D," in *Audio Engineering Society 141st Convention*, Los Angeles, 2016.

[2] K. Sunder, K. Jain, R. Cohen and G. Lurssen, "Reliable and Trustworthy Virtual Production Workflow for Surround and Atmos," in *Audio Engineering Society 153rd Convention*, 2022.

[3] K. Matsui and A. Ando, "Binaural Reproduction of 22.2 Multichannel Sound over Loudspeakers," in *Audio Engineering Society 129th Convention*, San Francisco, 2010.

[4] F. Rumsey, "Immersive audio Objects, mixing, and rendering," *Journal of Audio Engineering Society,* vol. 64, p. 584 – 588, 2016.

[5] A. De Sotgiu, M. Coccoli and G. Vercelli, "Comparing the perception of 'sense of presence' between a stereo mix and a binaural mix in immersive music," in *Audio Engineering Society 148th Convention*, 2020.

[6] R. L. Martin and K. I. McAnally, "Free-Field Equivalent Localization of Virtual Audio," *Journal of Audio Engineering Society,* vol. 49, pp. 14-22, 2001.

[7] J. C. Middlebrooks and D. M. Green, "Sound Localization by Human Listeners," *Annual Review of Psychology,* vol. 42, p. 135 – 159, 1991.

[8] J.-M. Pernaux, M. Emerit, J. Daniel and R. Nicol, "Perceptual Evaluation of Static Binaural Sound Synthesis," in *Audio Engineering Society 22nd International Conference on Virtual, Synthetic and Entertainment Audio*, 2002.

[9] M. T. Oramus and M. P. Neubauer, "Comparing Perception of Presence between Binaural and Stereo Reproduction," in *Audio Engineering Society International Conference: Audio for Virtual and Augmented Reality*, Redmond, 2022.

[10] G. Reardon, A. Genovese, G. Zalles, P. Flanagan and A. Roginska, "Evaluation of Binaural Renderers: Multidimensional Sound Quality Assessment," in *Audio Engineering Society Conference: Audio for Virtual and Augmented Reality*, Redmond, 2018.

[11] Y. Ueno, M. Mizumachi and T. Horiuchi, "Perceptual evaluation of binaural rendering and stereo width control in headphone reproduction," in *Audio Engineering Society 148th Convention*, 2020.

[12] G. Davidson, D. Darcy, L. Fielder, Z. Shuang, R. Graff, J. Breebaart and P. Crum, "Design and Subjective Evaluation of a Perceptually-Optimized Headphone Virtualizer," in *Audio Engineering Society 140th Convention*, Paris, 2016.

[13] G. Reardon, A. Genovese, G. Zalles, P. Flanagan and A. Roginska, "Evaluation of Binaural Renderers: Localization," in *Audio Engineering Society 144th Convention*, Milan, 2018.

[14] A. W. Mills, "On the Minimum Audible Angle," *Journal of the Acoustical Society of America,* vol. 30, no. 4, pp. 237-246, 1958.

[15] J. Blauert, "Binaural Room Simulation and Auditory Virtual Reality," in *Spatial Hearing: The Psychophysics of Human Sound Localization*, London, MIT Press, 1997, pp. 382-383.