Audio Engineering Society

Convention e-Brief 359

Presented at the 143rd Convention
2017 October 18–21, New York, NY, USA

Evaluation of Binaural Renderers: A Methodology

Gregory Reardon[1], Juan Simon Calle[2], Andrea Genovese[1], Gabriel Zalles[1], Marta Olko[1], Christal Jerez[1], Patrick Flanagan[2], and Agnieszka Roginska[1]

[1]*New York University*
[2]*THX*

Correspondence should be addressed to Gregory Reardon (gjr286@nyu.edu)

ABSTRACT

Recent developments in immersive audio technology have motivated a proliferation of binaural renderers used for creating spatial audio content. Binaural renderers leverage psychoacoustic features of human hearing to reproduce a 3D sound image over headphones. In this paper, a methodology for the comparative evaluation of different binaural renderers is presented. The methodological approach is threefold: a subjective evaluation of 1) quantitative characteristics (such as front/back and up/down discrimination and localization); 2) qualitative characteristics (such as naturalness and spaciousness); and 3) overall preference. The main objective of the methodology is to help to elucidate the most meaningful factors for the performance of binaural renderers and to provide insight on possible improvements in the rendering process.

1 Introduction

This paper focuses on establishing a procedure for the evaluation and characteristization of binaural technologies. Recent research in virtual reality (VR) and augmented reality (AR) has resulted in a growth of binaural technologies for rendering dynamic spatial audio. These processes leverage psychoacoustic features of human hearing to reproduce a 3D sound image over headphones [1, 2, 3]. When paired with head-tracking technologies, sound sources can be made to appear in a constant location with respect to the user's head orientation, creating the auditory illusion that sources are located in the same environment as the listener [4, 5, 6, 7, 8]. This is mandated for VR and AR technologies in order to create a coherent audiovisual image, thereby maintaining immersion and improving user presence [9, 10, 11, 12]. There is a need to create a standard procedure and set of metrics to use in the subjective evaluation of binaural headphone technologies, known in this paper as binaural renderers.

A comprehensive subjective methodology has been developed to judge the performance of a binaural renderer. The first phase of the test, known in this paper as the *quantitative* assessment, is focused on the fundamental evaluation of the 3D auditory image of the binaural renderer: externalization, front/back and up/down confusions, and localization. The second phase, known in this paper as the *qualitative* assessment, is focused on more general attributes of auditory image: naturalness, spaciousness, clarity, timbral balance, and dialogue intelligibility (movie stimuli only). The final phase consists of a *preference* assessment, where the user is able to rank the presented renderers from least preferred to most preferred. This ranking is later used to study correlations between sound quality attributes and listener preference. The terms *quantitative* and

qualitative used to describe the first two phases of the methodology serve only as identifiers. The attributes tested in the quantitative assessment are still sound quality attributes.

2 Background

Perceived sound quality has been shown to be comprised of distinct perceptual dimensions related to specific perceived sound properties [13]. Perceived sound quality can be expressed in terms of its overall value or as a function of its constituent perceptual characteristics. These two types of auditory assessment are known as global and parametric assessment respectively [14]. Letowski [14] presents a model for parametric assessment known as MURAL (Multilevel Auditory Assessment Language). In Letowski's model, overall sound quality is described as the multidimensional output of a hierarchical function of more specific sound qualities. Later authors cast doubt upon this particular partition of the sound quality space. Spaciousness, one of the two main subsets of sound quality for Letowski, has been differentiated from a number of other spatial attributes, such as spatial impression [15, 16]. Berg and Rumsey [17] keep a similar hierarchical structure, but replace *timbre* and *spaciousness*, the two main attribute classes for Letowski, with the following three main attribute classes: timbral, spatial, and technical.

Much work has been done attempting to define spatial attributes. Berg and Rumsey [18] employ a repertory grid technique (RGT) to elucidate these attributes. In the RGT, subjects first identify bipolar spatial sound attributes using their own vocabulary. Subjects then rate a set of stimuli using their own bipolar constructs. The cluster analysis of the constructs, presented in [19], yielded a number of significant spatial attributes. These spatial attributes, along with a set of other sound attributes, were tested in a follow-up experiment for multichannel loudspeaker reproduced sounds [20]. In the experiment, Berg and Rumsey broke up sound quality attributes tested into three classes: general, source, and room. The general attributes tested were naturalness, presence, preference, and envelopment. Source and room attributes included source width, localization, source distance, room width, and room size. Analysis revealed that subjects perceive the sound quality attributes as orthogonal along the dimensions of *general* and *source and room*. This breakup between general sound attributes and more specific source attributes informs the methodology proposed below. Other authors have attempted to identify significant spatial sound quality attributes using a number of techniques and over a number of reproduction systems [21, 22, 23, 24]. The set of sound quality attributes varies between authors. See [25] for a more comprehensive treatment of the elucidation methods for and definition of spatial sound quality attributes.

Binaural renderers seek to simulate the experience of a real or virtual acoustic environment. An assessment of sound quality reproduced by binaural renderers is dependent on the above sound quality attributes. The plausibility of the environment and the reproduction of virtual stimuli is dependent on how well the various constituent sound attributes, spatial, timbral, etc, can be made to be imperceptible from natural stimuli [12]. A number of quality features specific to binaural renderers have been identified and studied [26, 27, 28]. This work specifically focused on two main sound quality attributes: externalization and localization. Externalization refers to the perceived location of an auditory event either outside or within the head [29, 30, 31]. Various factors are known to affect externalization, the most significant of which are room divergence, which is an incongruence between synthesized scene and listening room, head tracking, and individualized head-related transfer functions (HRTFs). Localization refers to the perception of the correct direction of incidence of an auditory event. A particular type of localization error endemic in binaural renderers is that of reversal errors (front-back and up-down), which occur along auditory *cones of confusion*. Head tracking is known to reduce the number of front-back confusions [2, 8, 26].

3 Methodology

3.1 Quantitative Assessment

The first two phases of the methodology propose a parametric assessment of perceived sound quality. It is not feasible to test a comprehensive list of sound quality attributes, as discussed above, when evaluating multiple binaural renderers or even a single binaural renderer. Listener fatigue and time constraints limit the procedure to only a subset of sound quality attributes. This first phase of the methodology assesses externalization, front-back and up-down confusions, and localization. These can also be thought of as source specific sound qualities [20] as the stimuli presented are mono sources

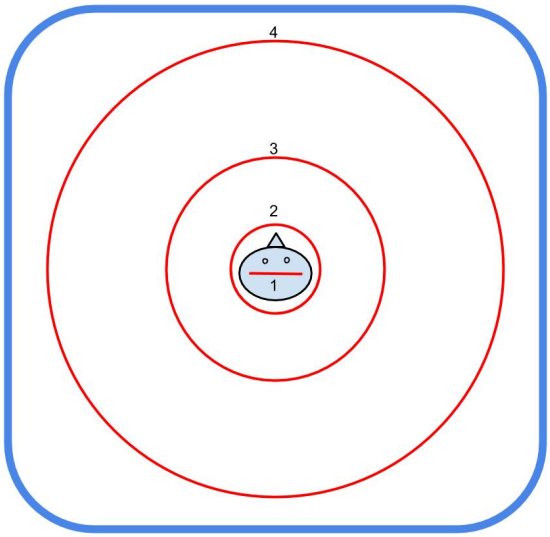

Fig. 1: Graphical representation of the discrete levels of externalization tested.

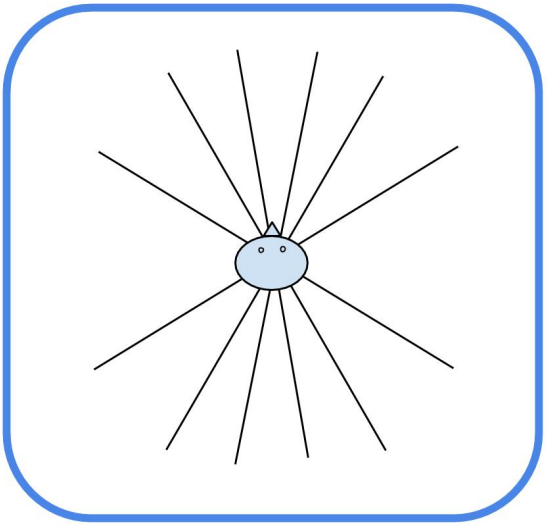

Fig. 2: Subdivisions of the localization regions on the horizontal plane.

virtualized and spatialized using a given binaural renderer. Though the binaural renderers of interest all support head-tracking for dynamic spatial audio on their native applications, the proposed methodology presents static spatial audio. Though this will have effects on measurements of externalization and reversal errors, it simplifies the experimental design and rendering experimental stimuli. Incorporating head tracking would also make it difficult to identify specific deficiencies at various azimuths and elevation, which might point to possible improvements in the rendering process.

In the externalization assessment, subjects are presented with a reference unprocessed stimulus, followed by three spatialized versions of the stimulus at random azimuths on the horizontal plane. The process is repeated for each renderer and each stimulus. Subjects are asked to rate the average level of externalization of the set of processed stimuli from one to four, where one represents "inside-the-head", and four is "far away from the head, external in space." An associated graphical representation of these levels of externalization accompanies the verbal descriptors and is pictured in *Fig 1*. Given that externalization is dependent upon the location of the sound source in space [26], a subset of equally spaced positions is randomly drawn from to determine the three spatialized stimuli to be presented. This requires gathering more participants but avoids any experimenter bias in selecting a subset of azimuth positions that might provide higher externalization on average. This returns an average value of externalization for a particular stimuli and given binaural renderer.

In the front-back portion of the test, subjects are presented with a pair of spatialized stimuli located along the same cone of confusion and asked to determine whether the trajectory of the pair was front-to-back or back-to-front. Similarly in the up-down portion of the test, subjects are presented with a pair of spatialized stimuli located at the same azimuth but with an elevation angle of either +30° or -30°. Subjects are asked to determine if the trajectory was up-to-down or down-to-up. Each trajectory was presented thrice in a row before asking for a response. The third and final part of the quantitative assessment consists of horizontal localization. Subjects are presented with a single spatialized stimuli in isolation and asked to determine the region from which the rendered audio source appeared to emanate. This graphic is pictured in *Fig 2*. Said regions are not equally spaced in order to reflect the resolution of human localization which is best suited at discriminating azimuth angles in the front and back regions, with resolution decreasing as the source moves towards the sides of the head [2].

3.2 Qualitative Assessment

The second phase of the methodology evaluated a set of general sound quality attributes. Virtual surround sound (VSS) stimuli were used. The attributes selected were Naturalness, Clarity, Spaciousness, Timbral Balance, and Dialogue Intelligibility. The length of the test once again limited the amount of general sound quality attributes that could be tested. The descriptions of each of the attributes is given:

- **Naturalness:** This attribute describes whether the sound gives a realistic impression, as opposed to artificial [32].
- **Clarity:** This attribute describes whether the sound appears to be clear or muffled [24].
- **Spaciousness:** This attribute describes how much the sound appears to surround you.
- **Timbral Balance:** This attribute describes how balanced (or colored) the different tone ranges of the sound appear to be.
- **Dialogue Intelligibility (movie stimuli only):** This attribute describes the ease at which dialogue can be understood.

Subjects were provided with a description of spaciousness closer to the typical definition of envelopment [32]. According to Rumsey, when subjects are surrounded by a group of dry sources in surround sound reproduction, envelopment is typically substituted into the listener's vocabulary [16]. Timbral balance was selected because the signal processing needed to produce a 3D sound image (ambisonics transformations, HRTF processing, etc.) requires coloring the spectrum to simulate natural interaural time and level differences [2]. Dialogue intelligibility was chosen to provide an additional measure for the movie stimuli that were tested.

Subjects were asked to discretely rate on a scale of 1 to 5 these attributes for surround sound clips rendered binaurally. The clips were presented side-by-side in lieu of a full-paired comparison test since there were a number of binaural renderers to be evaluated. No reference stereo downmix was included as VSS systems do not always perform as well as expected when compared against stereo down-mixed content [33, 34].

3.3 Preference Assessment

The third phase of the methodology is the global assessment of sound quality. In the preference assessment the same VSS stimuli used in the second phase are also used. Subjects are forced to select their *least* preferred clip from amongst the set of processed stimuli as it has often been found easier to ascertain one's least favorite clip [14]. The chosen clip is then removed and the remaining clips are presented again. The selection process continues until a complete ranking of the renderers is determined for each stimulus.

4 Conclusions

Given the commercial availability of a number of binaural renderers, there is a need and a desire to define a set of sound quality attributes for standard evaluation of these renderers. A background of sound quality assessment is first provided in section 2, examining both general sound quality attributes and those more relevant for the evaluation of binaural renderers. The proposed methodology motivated by this understanding is detailed in section 3. The methodology has three phases: quantitative, qualitative, and overall preference. The quantitative phase looks to assess externalization, front-back and up-down confusions, and general localization accuracy. The qualitative phase assesses *Naturalness*, *Spaciousness*, *Clarity*, *Timbral Balance*, and *Dialogue Intelligibility*, as detailed in section 3.2. The overall preference phase forces subjects to rank a set of stimuli from worst to best and allows for correlation studies to be performed in the subsequent data analysis. Testing is currently underway and results of the methodology, along with more specific details of an experiment, are to be presented in a future publication.

References

[1] Kleiner, M., Dalenbäck, B.-I., and Svensson, P., "Auralization-an overview," *Journal of the Audio Engineering Society*, 41(11), pp. 861–875, 1993.

[2] Blauert, J., *Spatial hearing: the psychophysics of human sound localization*, MIT press, 1997.

[3] Begault, D. R. and Trejo, L. J., "3-D sound for virtual reality and multimedia," 2000.

[4] Wallach, H., "The role of head movements and vestibular and visual cues in sound localization." *Journal of Experimental Psychology*, 27(4), p. 339, 1940.

[5] Thurlow, W. R. and Runge, P. S., "Effect of induced head movements on localization of direction of sounds," *The Journal of the Acoustical Society of America*, 42(2), pp. 480–488, 1967.

[6] Wightman, F. L. and Kistler, D. J., "The importance of head movements for localizing virtual auditory display objects," Georgia Institute of Technology, 1994.

[7] Wenzel, E. M., "Effect of increasing system latency on localization of virtual sounds," in *Audio Engineering Society Conference: 16th International Conference: Spatial Sound Reproduction*, Audio Engineering Society, 1999.

[8] Begault, D. R., Wenzel, E. M., and Anderson, M. R., "Direct comparison of the impact of head tracking, reverberation, and individualized head-related transfer functions on the spatial perception of a virtual speech source," *Journal of the Audio Engineering Society*, 49(10), pp. 904–916, 2001.

[9] Wenzel, E. M., Fisher, S. S., Stone, P. K., and Foster, S. H., "A system for three-dimensional acoustic visualization in a virtual environment workstation," in *Proceedings of the 1st conference on Visualization'90*, pp. 329–337, IEEE Computer Society Press, 1990.

[10] Begault, D. R., Ellis, S. R., and Wenzel, E. M., "Headphone and head-mounted visual displays for virtual environments," in *Audio Engineering Society Conference: 15th International Conference: Audio, Acoustics & Small Spaces*, Audio Engineering Society, 1998.

[11] Schubert, T., Friedmann, F., and Regenbrecht, H., "The experience of presence: Factor analytic insights," *Presence: Teleoperators and virtual environments*, 10(3), pp. 266–281, 2001.

[12] Lindau, A. and Weinzierl, S., "Assessing the plausibility of virtual acoustic environments," *Acta Acustica united with Acustica*, 98(5), pp. 804–810, 2012.

[13] Gabrielsson, A., "Dimension analyses of perceived sound quality of sound-reproducing systems," *Scandinavian Journal of Psychology*, 20(1), pp. 159–169, 1979.

[14] Letowski, T., "Sound quality assessment: concepts and criteria," in *Audio Engineering Society Convention 87*, Audio Engineering Society, 1989.

[15] Griesinger, D., "The psychoacoustics of apparent source width, spaciousness and envelopment in performance spaces," *Acta Acustica united with Acustica*, 83(4), pp. 721–731, 1997.

[16] Rumsey, F., "Spatial quality evaluation for reproduced sound: Terminology, meaning, and a scene-based paradigm," *Journal of the Audio Engineering Society*, 50(9), pp. 651–666, 2002.

[17] Berg, J. and Rumsey, F., "Systematic evaluation of perceived spatial quality," in *Audio Engineering Society Conference: 24th International Conference: Multichannel Audio, The New Reality*, Audio Engineering Society, 2003.

[18] Berg, J. and Rumsey, F., "Spatial attribute identification and scaling by repertory grid technique and other methods," in *Audio Engineering Society Conference: 16th International Conference: Spatial Sound Reproduction*, Audio Engineering Society, 1999.

[19] Berg, J. and Rumsey, F., "In search of the spatial dimensions of reproduced sound: Verbal protocol analysis and cluster analysis of scaled verbal descriptors," in *Audio Engineering Society Convention 108*, Audio Engineering Society, 2000.

[20] Rumsey, F. and Berg, J., "Verification and correlation of attributes used for describing the spatial quality of reproduced sound," in *Audio Engineering Society Conference: 19th International Conference: Surround Sound-Techniques, Technology, and Perception*, Audio Engineering Society, 2001.

[21] Zacharov, N. and Koivuniemi, K., "Unravelling the perception of spatial sound reproduction: Language development, verbal protocol analysis and listener training," in *Audio Engineering Society Convention 111*, Audio Engineering Society, 2001.

[22] Guastavino, C. and Katz, B. F., "Perceptual evaluation of multi-dimensional spatial audio reproduction," *The Journal of the Acoustical Society of America*, 116(2), pp. 1105–1115, 2004.

[23] Choisel, S. and Wickelmaier, F., "Extraction of auditory features and elicitation of attributes for the assessment of multichannel reproduced sound," *Journal of the Audio Engineering Society*, 54(9), pp. 815–826, 2006.

[24] Lorho, G., "Evaluation of spatial enhancement systems for stereo headphone reproduction by preference and attribute rating," in *Audio Engineering Society Convention 118*, Audio Engineering Society, 2005.

[25] Le Bagousse, S., Colomes, C., and Paquier, M., "State of the art on subjective assessment of spatial sound quality," in *Audio Engineering Society Conference: 38th International Conference: Sound Quality Evaluation*, Audio Engineering Society, 2010.

[26] Werner, S. and Klein, F., "Influence of Context Dependent Quality Parameters on the Perception of Externalization and Direction of an Auditory Event," in *Audio Engineering Society Conference: 55th International Conference: Spatial Audio*, Audio Engineering Society, 2014.

[27] Werner, S., Klein, F., Mayenfels, T., and Brandenburg, K., "A summary on acoustic room divergence and its effect on externalization of auditory events," in *Quality of Multimedia Experience (QoMEX), 2016 Eighth International Conference on*, pp. 1–6, IEEE, 2016.

[28] Werner, S., Götz, G., and Klein, F., "Influence of Head Tracking on the Externalization of Auditory Events at Divergence between Synthesized and Listening Room Using a Binaural Headphone System," in *Audio Engineering Society Convention 142*, Audio Engineering Society, 2017.

[29] Mills, A. W., "Lateralization of High-Frequency Tones," *The Journal of the Acoustical Society of America*, 32(1), pp. 132–134, 1960.

[30] Plenge, G., "On the problem of "in head localization"," *Acta Acustica united with Acustica*, 26(5), pp. 241–252, 1972.

[31] Hartmann, W. M. and Wittenberg, A., "On the externalization of sound images," *The Journal of the Acoustical Society of America*, 99(6), pp. 3678–3688, 1996.

[32] Choisel, S. and Wickelmaier, F., "Evaluation of multichannel reproduced sound: Scaling auditory attributes underlying listener preference," *The Journal of the Acoustical Society of America*, 121(1), pp. 388–400, 2007.

[33] Zacharov, N. and Lorho, G., "Subjective evaluation of virtual home theatre sound systems for loudspeakers and headphones," in *Audio Engineering Society Convention 116*, Audio Engineering Society, 2004.

[34] Pike, C. and Melchior, F., "An assessment of virtual surround sound systems for headphone listening of 5.1 multichannel audio," in *Audio Engineering Society Convention 134*, Audio Engineering Society, 2013.