Audio Engineering Society

Convention Paper 10272

Presented at the AES 160th Convention
2026 May 28–30, Copenhagen, Denmark

This paper was peer-reviewed as a complete manuscript for presentation at this convention. 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.

Perceptual Evaluation of the Open Binaural Renderer

Katarzyna Sochaczewska[1], Tomasz Rudzki[1], Gavin Kearney[1], Jan Skoglund[2], Felicia S. C. Lim[2], and Jani Huoponen[2]

[1]*University of York, York, United Kingdom*
[2]*Google LLC, Mountain View, California, USA*

Correspondence should be addressed to Katarzyna Sochaczewska (k.sochaczewska@york.ac.uk)

ABSTRACT

This paper presents the perceptual evaluation of the Open Binaural Renderer (OBR), an open-source library developed for headphone-based rendering of Immersive Audio Model and Formats (IAMF) content. The evaluation followed an iterative framework in which findings from a pilot listening study informed the tuning of rendering profiles, and the resulting renderer was benchmarked against established proprietary solutions. In the pilot study, 19 expert listeners rated the Overall Listening Experience (OLE) of the initial prototype (OBR_{v1}) and five external renderers across diverse audio content. Qualitative feedback was analysed using inductive coding to identify salient perceptual dimensions. The pilot revealed content-dependent performance and showed that a single default profile was inadequate, yielding mixed responses in both the numerical scale and in the qualitative feedback and motivating the development of multiple rendering profiles in OBR_{v2}. The main study evaluated two OBR_{v2} profiles targeting different reverberation characteristics (*Direct* and *Ambient*) alongside three top-performing external renderers. A total of 39 participants, divided into expert and non-expert groups, rated five perceptual attributes: Voice Quality, Envelopment, Externalisation, Overall Listening Experience, and Timbral Balance. Mixed-design ANOVA revealed significant main effects of renderer condition on all attributes. Pairwise comparisons showed that $OBR_{v2,Ambient}$ achieved significantly higher OLE ratings than one proprietary renderer and reached statistical parity with the remaining two, representing a measurable improvement over the prototype. A trade-off between Voice Quality and Externalisation was observed, driven by the level of reverberation in each renderer. The results demonstrate that iterative, perceptually informed tuning can yield competitive binaural rendering quality in an open-source framework.

1 Introduction

Binaural rendering of immersive audio content is at the center of the media streaming revolution. It is used to deliver immersive audio experiences over headphones by simulating the listening conditions of an acoustically treated room equipped with a multichannel speaker array. Due to its hardware simplicity, headphone-based listening is significantly more accessible than using a multichannel loudspeaker system in a home environment or a purpose-built listening room. While major video and music streaming platforms adopt interactive, head-tracked binaural rendering using their proprietary solutions, the availability of free and open-source software utilizing state-of-the-art rendering algorithms is limited. Immersive Audio Model and Formats (IAMF) [1], also known as Eclipsa Audio, is an open-source audio container developed to accom-

modate multichannel and scene-based audio formats. Headphone-based delivery of IAMF audio requires efficient binaural rendering, and therefore the Open Binaural Renderer (OBR) was developed specifically to render IAMF audio content.

The OBR is an open-source initiative maintained through an ongoing process of iterative evaluation and refinement. Its architecture [2] renders channel-based, object-based, and Ambisonic content via an intermediate spherical-harmonic representation of head-related impulse responses (HRIRs), incorporating MagLS [3] binaural filters and a hybrid room model that combines anechoic HRIRs with a listening-room response. Three purposely tuned profiles are provided: *Direct* (anechoic), *Ambient* (moderate room reflections), and *Reverberant* (enhanced spaciousness), each supporting Ambisonic orders from 1st to 4th. An iterative evaluation framework and objective analysis were presented in [4], where assessment metrics actively drove the development and tuning process. This paper presents the perceptual evaluation of OBR, detailing how findings from a pilot study informed the tuning of its binaural rendering profiles and how the resulting renderer was benchmarked against established proprietary solutions through a multi-attribute listening test.

2 Evaluation Framework

OBR was benchmarked against anonymized existing renderers through both objective and perceptual evaluation. The objective results, including frequency responses, Interaural Level Difference (ILD), Interaural Time Difference (ITD), Interaural Cross-Correlation Coefficient (IACC), Direct-to-Reverberant Ratio (DRR), Reverberation Time (RT_{60}), and Interaural Coherence (IC), were reported in [4]. The perceptual evaluation is the focus of this paper.

2.1 Evaluated Renderers

The selection of external renderers for evaluation was governed by three criteria: (1) integration of both Channel-Based (CBA) and Object-Based (OBA) workflows; (2) compatibility with the Audio Definition Model (ADM) [5] format; and (3) established adoption within the professional music industry. These criteria defined the renderer pool for the pilot study. For the main study, only the top-performing external renderers (highest OLE scores in the pilot) were retained, and two tuned OBR_{v2} profiles were added. Table 1 summarises the renderer conditions included in each study.

OBR utilises anechoic Head-Related Transfer Functions (HRTFs) of the Neumann KU100 dummy head from the SADIE II database [6], while $Renderer_B$ employs Binaural Room Impulse Responses (BRIRs) measured with a Neumann KU100 in an ITU-R BS.1116-compliant listening room. Information about the HRTFs used in the remaining renderers is not publicly available. In the pilot study, OBR employed 7th-order Ambisonic binaural filters. For the main study, the filter order was reduced to 3rd-order following findings from an intermediate investigation; full details of this change will be reported in an upcoming extended study.

Renderer	Pilot	Main Study
A	✓	✓
B	✓	
C	✓	✓
D	✓	✓
E	✓	
OBR_{v1}	✓	
OBR_{v1EQ}	✓	
$OBR_{v2,Direct}$		✓
$OBR_{v2,Ambient}$		✓
Stereo	✓	

Table 1: Renderers included in the pilot and the main studies.

2.2 Common Methodology

The testing protocol focused solely on binaural comparison without a loudspeaker reproduction reference. Including a loudspeaker setup would have provided a more comprehensive evaluation. However, it would also exclude remote participants, restricting the pool of available experts and substantially complicating the test design, since switching between headphone and loudspeaker playback is rarely seamless. Both the pilot study and the main study employed a MUSHRA-style interface for direct assessment of perceptual attributes on a continuous scale, without a designated reference signal. Renderers were anonymized across both studies. All audio samples were standardized to 30 s excerpts sourced from 7.1.4 channel-based masters. The Overall Listening Experience (OLE) served as the primary global quality metric throughout the evaluation process. In the pilot study, OLE was rated on a 0–100 scale; in

the main study, all attributes including OLE were rated on a bipolar −50 to +50 scale.

3 Pilot Study & Informed Tuning

Assessing binaural rendering quality requires deconstructing the multidimensional nature of auditory perception [7] into quantifiable components via both global and parametric assessments [8]. A pilot listening study was conducted with two goals: (a) identify perceptual deficiencies of OBR_{v1}, and (b) elicit salient attributes to inform both subsequent tuning and the main study design. Eight conditions were evaluated (see Table 1): OBR_{v1}; OBR_{v1EQ}, which incorporated a preliminary spectral correction curve aimed at increasing clarity, achieved directly via a high-frequency shelf (>2000 Hz) and indirectly by attenuating the 250 Hz midrange region; a popular open-source renderer; two proprietary renderers, one of which was tested in three binaural modes; and a stereo fold-down.

3.1 Pilot Design

Following the common methodology described in Section 2, the pilot test additionally included an open feedback form to capture qualitative impressions. The study included 19 subjects (15 male, 4 female; median age = 37 years, sd = 9.2 years). Five audio samples were selected to cover diverse scenes, ensuring a mix of music, speech, and sound effects. These included a *Classical* recording and a *Score* recording from the ECHO Project [9], captured using the 2L Prism immersive microphone array, as well as *Podcast* (Documentary), *Storytelling* (Audiobook), and *Folk* (in 5.0 format). The study evaluated binaural profiles in isolation rather than within the context of productions which blend different profiles together.

3.2 Pilot Results

The results of the pilot study (Fig. 1) confirmed that the single default profile of OBR_{v1} was inadequate for the wide range of audio content encountered in practice, yielding mixed responses for different audio samples in both the numerical scale and in the qualitative feedback. Perceptual evaluation clearly demonstrated content-dependent performance, with OLE scores varying substantially across different genres and sample types. This finding, coupled with evidence that universal spectral correction was ineffective, led to the development of discrete rendering profiles optimised for specific content characteristics in OBR_{v2}.

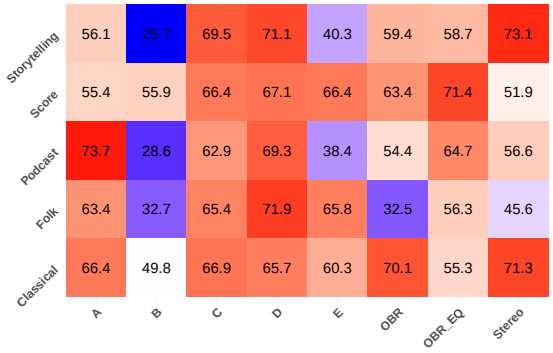


Fig. 1: Mean OLE scores of the examined renderers for different audio content and context.

In addition to quantitative ratings, qualitative feedback provided further insight into the perceptual dimensions underlying these scores. By restricting the preliminary study to expert subjects in accordance with ITU-R BS.1534-3, the resulting qualitative feedback had sufficient semantic consistency to support inductive coding [10]. The initial codes were grouped into categories, presented in Fig. 2. Attributes marked with an asterisk were always directly proportional: a negative judgement denoted an unwanted quality, while a positive judgement denoted a preferred quality. Other categories were subject to preference; for instance, a "natural" sound could be followed by "boring", and high externalisation could be followed by "artificial" or "weird".

3.3 Informed Tuning

The quantitative and qualitative findings from the pilot study, combined with the literature review [6, 8, 11], helped support the selection of attributes for the main study and identify areas for refinement, which were addressed in OBR_{v2}. Table 2 lists the main observations and how they were addressed in OBR_{v2}. Together, these observations informed both the tuning of OBR_{v2} profiles and the selection of attributes and stimuli for the main study described in Section 4.

4 Main Study

Building on the pilot findings, the main study made three key design changes. First, the qualitative analysis

Table 2: Pilot study observations and corresponding actions taken.

Observation	Action Taken
The OBR_{v1} lacked clarity and high-frequency definition. Introducing a spectral equalisation curve on the rendered binaural output resulted in improvement in OLE over OBR without EQ in some examples, indicating that spectral correction can be beneficial.	More informed tuning was performed: an equalization (high shelf curves) was introduced and applied separately to anechoic and reverberated parts.
The voice quality was revealed as a salient perceptual attribute, which appeared to correlate with both DRR implemented in each renderer and the contextual expectations associated with different content types. Specifically, renderers with higher reverberation levels (lower DRR values) received lower ratings for speech-centric content, while the same characteristics were less detrimental to musical material. The preference for added reverberation appeared to depend on the audio context. However, using a default virtual room that most likely diverges from the listener's own, negatively affects critical listening. At the same time, reverberation positively influences the externalisation of sounds such as noise and speech [12, 13], but as observed, it can also degrade Voice Quality.	Multiple profiles (Direct, Ambient, Reverberant) with varied level of reverberation were introduced. The detailed description of the profiles is outside of the scope of this study and will be a focus of a separate publication.
More frequent mentions of envelopment for the OBR_{v1EQ} condition might suggest that the enhanced high-frequency details may contribute positively to this attribute.	Included *Envelopment* as one of the attributes to further investigate in the main study.
Complex audio scenes, such as orchestral recordings or samples with transients, such as percussive patterns, were often described as difficult to judge and/or express preference. The ratings for *Score* and *Classical* samples revealed the least differences in ratings across all renderers (see Fig. 1).	The *Score* sample was removed from the audio samples set; the *Orchestral* sample was retained for timbral fidelity evaluation.
OLE scores reflect strongly subjective judgements and produced contrasting opinions on certain conditions. Some qualities with seemingly positive connotation, such as natural sound, are not always preferred. Some participants coupled "Natural" with "Boring" or "Not Exciting"; strong content dependence further influences this judgement.	Future investigation
Reduced low-end is usually paired with better spaciousness and externalisation, but compromises on "Fullness" and general preference. Timbral Balance in audio samples is crucial for adequate judgement.	Future investigation

revealing distinct perceptual dimensions (Section 3) motivated a shift from global to parametric evaluation. Second, the renderer pool was narrowed based on OLE rankings: the stereo fold-down, OBR_{v1}, OBR_{v1EQ}, and two external renderers with the lowest pilot scores were removed. Third, two new OBR_{v2} profiles, a direct result of the informed tuning process (Section 3.3), were added, yielding five conditions in total (see Table 1).

4.1 Perceptual Attributes

Thematic analysis of the pilot qualitative feedback (Section 3), synthesised with existing literature, identified recurring perceptual dimensions. Consequently, the main evaluation assessed five discrete attributes encompassing both timbral and spatial qualities (see Table 3).

4.2 Listening Test Design

The main evaluation retained the three top-performing external renderers from the pilot study alongside two OBR_{v2} profiles: $OBR_{v2,Ambient}$ and $OBR_{v2,Direct}$.[1] The ratings were collected on a continuous bipolar scale (-50 to $+50$) where higher values denoted superior

[1] A third proposed profile, $OBR_{v2,Reverberant}$, was excluded from this evaluation; it is not expected to be used in isolation.

Table 3: Attributes with definitions and end-points

Attribute	Definition	Low −50	Mid 0	High +50
Voice Quality	Evaluate the sound stimuli in terms of voice quality using your subjective judgement.	Bad	Fair	Excellent
Envelopment	Degree of being surrounded by a source, scene or ensemble / Feeling of being located in the middle of the audio scene	Not enveloping	Neither enveloping nor not-enveloping	Completely enveloping
Externalisation	Does the sound appear to be inside your head or outside of your head?	Inside of the head	Neither outside nor inside of the head	Outside of the head
Overall Listening Experience	Evaluate the sound stimuli in terms of your preference	Strongly dislike	Neither like nor dislike	Strongly like
Timbral Balance	Denotes the balance between bass and treble. Dark: Excessive bass; Neutral: Bass and treble perceived equally loud; Bright: Excessive treble.	Dark	Neutral (well-balanced)	Bright

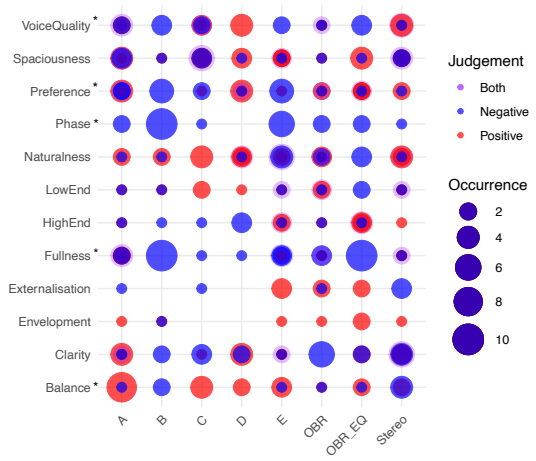


Fig. 2: Qualitative analysis of pilot study feedback across all audio samples.

performance, with the exception of *Timbral Balance*, which employed a centered optimum at 0. To ensure statistical consistency with other metrics, the latter scores were subsequently transformed for analysis:

$$S^* = 50 - |S_{raw}| \tag{1}$$

where S^* is the transformed score and S_{raw} is the original Timbral Balance rating.

Timbral fidelity has been shown to constitute up to 70% of the judgment of the audio quality [14]. The authors acknowledge the importance of timbral fidelity as well as limitations inherent in this simplified assessment of timbral balance [15]; a more comprehensive approach would independently evaluate low, mid, and high frequency ranges; however, such granularity would have extended the duration and cognitive load beyond acceptable limits for the current study design. Future work will include a more extensive timbral fidelity evaluation focused specifically on musical content.

4.3 Test Stimuli

To evaluate the perception of timbral and spatial fidelity, stimuli were curated to encompass a broad range of contexts, involving samples of complex spectral character, with both static and dynamic spatial content. The dataset retained three core excerpts from the pilot study (Orchestral, Documentary, and Audiobook) to ensure continuity, while expanding the scope with additional material including electronic music, broadcast speech, and dynamic sound design. Collectively, these stimuli featured varying degrees of reverberation and spatial complexity (e.g., static true center, phantom center, stereo pairs, and dynamic trajectories). As in the pilot study, all samples were standardized to a duration of 30 s and sourced from 7.1.4 channel-based masters. For the full list and description of stimuli, see Table 4.

4.4 Test Participants

Prior research indicated that listener training and familiarity with stimuli significantly enhance the ability

Table 4: Audio samples included in the test. Asterisk marks samples retained from the pilot study.

Sample	Description
Orchestral* (music)	Orchestral recording from ECHO Project [9], captured with 2L-Prism microphone array.
Documentary* (speech + sound effects)	Female speech (English language, phantom center), room ambiance, environmental Foley (e.g., crackling fire, ticking clock) distributed around the scene; custom production.
Audiobook* (speech + music)	*The Dogs* by Michael Stewart; male voice-over (English language, true center) with immersive sound design; custom production, courtesy of the author.
Broadcast (speech)	Female speech (Polish language, true center); dry recording using a dynamic microphone.
Podcast (speech)	Two male voices (English/German languages, stereo pair L/R); condenser microphone recording with room ambiance.
Audioplay (speech + sound design)	Female voice (English language) with automated trajectory (circular movement in the horizontal plane) and accompanying sound effects.
Pop/Soul (music with vocals)	Music with female vocal track; 7.1.4 remix from stems from initially stereo production.
Electronic (music)	Original electronic composition produced specifically for immersive format; custom production.

to evaluate spatial and timbral attributes [14]. Consequently, participation was restricted to individuals demonstrating relevant background experience, specifically: spatial audio researchers, musicians, and recording or mixing engineers who had prior experience with multichannel workflows.

A total of 39 participants were divided into two groups based on their professional experience. The *Experts* group comprised individuals with over 15 years of experience or an established career in audio engineering or spatial audio ($N = 20$; median age 39 years, sd 9). The *Non-Experts* group comprised participants with limited experience in sound engineering, spatial audio or music ($N = 19$; median age 30 years, sd 9.4). The gender distribution included 8 females (4 per group) and 31 males (16 expert, 15 non-expert, Fig. 3). Sixteen of the participants were also part of the listening panel in the pilot study prior to the main test. All subjects reported normal hearing. Regarding playback hardware, sixteen subjects used Sennheiser HD650 headphones, qualified as *reference headphones*, as they were used for all renderer tuning. The remaining participants used various other studio-grade models (Tab. 5). There was insufficient data for any other single model to support a meaningful analysis of the headphone effect, although it is acknowledged as a potential source of variability. When pooled into an "other headphones" group, no statistically significant differences were observed.

Table 5: Headphone models used in test by participants and their frequency of occurrence ($N = 39$).

Model	Occurrence
AKG K712	1
Audeze LCD-X	1
Audeze MM-500	1
Audio-Technica ATH-M50x	3
Beyerdynamic DT770 (incl. Pro)	3
Beyerdynamic DT880	1
Beyerdynamic DT990 Pro	1
Focal (unspecified)	1
Neumann NDH30	1
Sennheiser HD280 Pro	1
Sennheiser HD560s	1
Sennheiser HD600	3
Sennheiser HD650	16
Sennheiser HD800	1
Sennheiser Momentum	1
Sony MDR-7506	1
Sony MDR-MV1	2

Subjects were compensated with vouchers valued at £20 (or equivalent in local currency). The

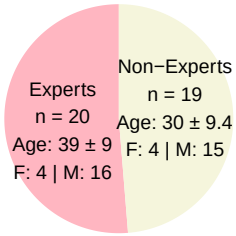


Fig. 3: Demographics data including median age with standard deviation and gender distribution.

study complied with University of York's Physical Sciences Ethics Committee (ethics consent number Sochaczewska20250321).

5 Results

Data analysis was conducted at both the population level and stratified by experience. The results for those two groups and all of the investigated attributes are presented in Fig. 4. The Expert group demonstrated greater criticality in the assessment of *Envelopment*, yielding consistently lower mean scores than Non-Experts. Conversely, Experts rated the *Externalisation* of $OBR_{v2,Ambient}$ significantly higher. No statistically significant inter-group differences were observed for the remaining attributes.

The sets of answers from three participants are not complete due to connectivity issues of the testing platform. Data were first checked for outliers and normality. Shapiro-Wilk tests indicated deviations from normality in specific conditions ($p < 0.05$). While sporadic outliers were detected, they were deemed to represent genuine variations in subjective response rather than measurement artifacts. Consequently, all data points were retained. Sensitivity checks confirmed that the inclusion of extreme outliers did not alter the statistical results. Sphericity was assessed using Mauchly's test, with Greenhouse-Geisser corrections applied where necessary. The mixed-design ANOVA revealed a significant main effect of the Renderer Condition on all attribute ratings (see Table 6).

Renderers which do not introduce reverberation generally yielded superior ratings for *Voice Quality* at the expense of *Externalisation* and *Envelopment*. Regarding timbral fidelity, $OBR_{v2,Ambient}$ demonstrated greater spectral neutrality (lower coloration)

Table 6: Mixed ANOVA results for Renderer as the main effect.

Attribute	F	p	η^2_G
Envelopment	3.938	0.010	0.042
Externalisation	15.145	$< .001$	0.163
Timbral Balance	6.045	$< .001$	0.067
Voice Quality	19.662	$< .001$	0.226
OLE	3.980	0.004	0.048

than both $Renderer_A$ and $Renderer_D$. Furthermore, $OBR_{v2,Ambient}$ achieved a significantly higher OLE rating compared to $Renderer_A$. All statistically significant pairwise comparisons are summarised in Table 7 (reporting Bonferroni-adjusted p-values).

The restriction of specific stimuli to relevant attributes (e.g., speech samples for *Voice Quality*) resulted in a non-uniform dataset, constraining the cross-correlation analysis. Within this limited set, Pearson coefficients indicated a high degree of independence between perceptual dimensions. *Voice Quality* showed the strongest association with *OLE* ($r = 0.39$), suggesting a moderate predictive relationship, while correlations for the remaining attributes were negligible ($r < 0.15$). Aside from a weak relationship between *Envelopment* and *Externalisation* ($r = 0.25$), the low inter-attribute covariance suggests that *OLE* was not dominated by any single spatial or timbral parameter.

6 Discussion

This study presented a methodology for the perceptual evaluation of binaural rendering of channel-based audio content, applied to the iterative development of OBR. A pilot study informed the tuning of rendering profiles and identified critical perceptual dimensions; the main study then benchmarked two optimised OBR_{v2} profiles against three external renderers.

Potential familiarity bias was not explicitly controlled in this study. It is reasonable to assume that expert participants possessed prior exposure to the proprietary renderers or their underlying HRTFs. Deviations from such reference conditions often result in lower initial preference ratings, a bias known to diminish with exposure and training [16]. Future work could isolate this variable via a parametric study focused on long-term adaptation.

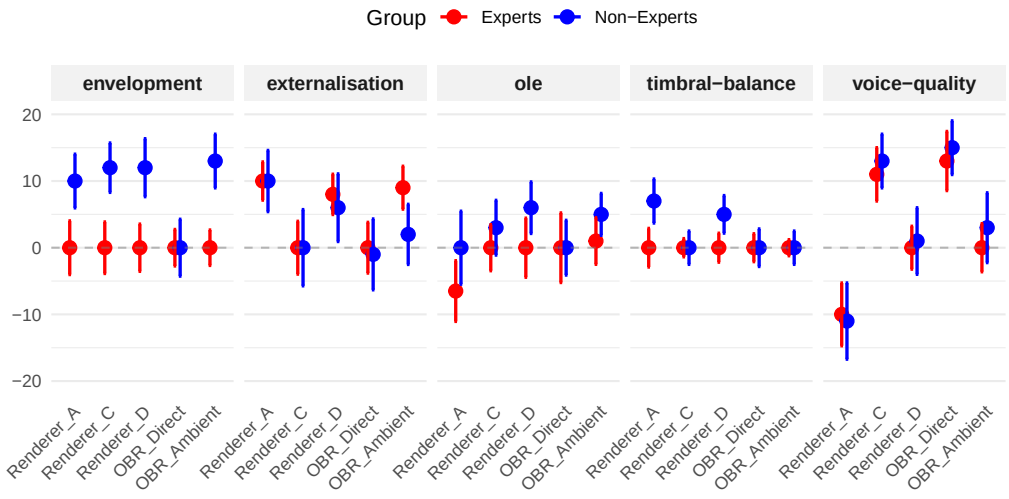


Fig. 4: Rating Scores for Experts and Non-Experts. Medians and associated non-parametric 95% confidence intervals for data obtained from the listening tests (notch edges are removed for readability). The values for Timbral Balance are not transformed here (0 is an optimal value).

The integration of pilot feedback into the tuning processes yielded a measurable improvement in the overall listening experience. In the pilot study, OBR_{v1} was rated significantly lower than the reference systems. Conversely, in the follow-up study, the optimised $OBR_{v2,Ambient}$ profile performed significantly better than $Renderer_A$. Furthermore, comparisons against $Renderer_C$ and $Renderer_D$ shifted from a significant deficit in the pilot to statistical parity (no significant difference) in this iteration. While OBR_{v2} currently employs a generalised ("non-individualised") processing approach, the introduction of selectable rendering profiles offers a mechanism to further capitalise on listener preference, particularly by addressing the emerging preference for anechoic speech rendering via *Voice Quality* ratings.

The variance in *Timbral Balance* ratings may reflect the limitations of a unidimensional bipolar scale (Bright vs. Dark). Expert listeners in some examples could hear prominent low and high frequencies at the same time, and the inability to rate each band independently potentially confounded the ratings for this group more than the Non-Experts group.

The authors acknowledge the multidimensional aspects of binaural rendering and associated confounding variables that influence perception and preference. Further work is needed in this realm, with focus on translation between speaker playback systems, the virtual room divergence problem, and different content types and formats, both channel- and scene-based.

7 Conclusion

This paper presented the perceptual evaluation of the Open Binaural Renderer as part of an iterative development process. A pilot study identified content-dependent performance limitations of the initial prototype and elicited salient perceptual attributes, which informed the design of multiple rendering profiles in OBR_{v2}. The subsequent main study benchmarked two OBR_{v2} profiles against three proprietary renderers across five perceptual attributes. Results showed that $OBR_{v2,Ambient}$ achieved significantly higher OLE ratings than $Renderer_A$ and reached statistical parity with $Renderer_C$ and $Renderer_D$, demonstrating measurable improvement over the prototype. The evaluation also revealed a trade-off between voice quality and externalisation, driven by the level of reverberation in each renderer. Future work will extend the evaluation to scene-based and object-based content and address timbral fidelity assessment with greater granularity. A reference-based study involving OBR comparison with loudspeaker playback is also planned.

Table 7: Pairwise comparisons of mean scores across renderers. Only statistically significant differences are listed. The higher scoring renderer is highlighted in **bold**.

Group 1	Group 2	n_1	n_2	t-stat	df	p_{adj}
Attribute: Voice Quality						
$Renderer_A$	$\mathbf{Renderer_C}$	39	39	−6.12	38	< 0.01
$Renderer_A$	$\mathbf{Renderer_D}$	39	39	−4.61	38	< 0.01
$Renderer_A$	$\mathbf{OBR_{v2,Direct}}$	39	39	−5.76	38	< 0.01
$Renderer_A$	$\mathbf{OBR_{v2,Ambient}}$	39	39	−4.60	38	< 0.01
$\mathbf{Renderer_C}$	$Renderer_D$	39	39	3.76	38	< 0.01
$\mathbf{Renderer_C}$	$OBR_{v2,Ambient}$	39	39	3.05	38	0.042
$Renderer_D$	$\mathbf{OBR_{v2,Direct}}$	39	39	−3.64	38	< 0.01
$\mathbf{OBR_{v2,Direct}}$	$OBR_{v2,Ambient}$	39	39	3.11	38	0.035
Attribute: Envelopment						
$OBR_{v2,Direct}$	$\mathbf{OBR_{v2,Ambient}}$	37	37	−3.71	36	< 0.01
Attribute: Externalisation						
$\mathbf{Renderer_A}$	$Renderer_C$	38	38	3.72	37	< 0.01
$\mathbf{Renderer_A}$	$OBR_{v2,Direct}$	38	38	5.00	37	< 0.01
$Renderer_C$	$\mathbf{Renderer_D}$	38	38	−3.64	37	< 0.01
$\mathbf{Renderer_D}$	$OBR_{v2,Direct}$	38	38	5.66	37	< 0.01
$OBR_{v2,Direct}$	$\mathbf{OBR_{v2,Ambient}}$	38	38	−5.10	37	< 0.01
Attribute: Timbral Balance						
$Renderer_A$	$\mathbf{Renderer_C}$	38	38	−3.38	37	0.017
$Renderer_A$	$\mathbf{OBR_{v2,Ambient}}$	38	38	−4.35	37	< 0.01
$Renderer_D$	$\mathbf{OBR_{v2,Ambient}}$	38	38	−3.45	37	0.014
Attribute: OLE						
$Renderer_A$	$\mathbf{Renderer_D}$	39	39	−3.53	38	0.011
$Renderer_A$	$\mathbf{OBR_{v2,Ambient}}$	39	39	−3.72	38	< 0.01

Acknowledgment

The authors would like to thank the participants for taking part in the study. This research is supported by Google.

References

[1] Alliance for Open Media, "Immersive Audio Model and Formats," Specification v1.1.0, Alliance for Open Media, 2024.

[2] Rudzki, T., Kearney, G., and Skoglund, J., "On the Design of the Binaural Rendering Library for Eclipsa Audio Immersive Audio Container," in *158th Convention of the Audio Engineering Society*, Warszawa, Poland, 2025.

[3] Schörkhuber, C., Zaunschirm, M., and Höldrich, R., "Binaural rendering of ambisonic signals via magnitude least squares," in *Proceedings of the DAGA*, volume 44, pp. 339–342, 2018.

[4] Sochaczewska, K., Rudzki, T., Kearney, G., and Skoglund, J., "Eclipsa Audio – Open Binaural Renderer: Objective and Perceptual Evaluation," in *Tonmeistertagung Proceedings*, Tonmeistertagung, 2025.

[5] International Telecommunication Union, "Recommendation ITU-R BS.2076-3: Audio definition model," Recommendation BS.2076-3, International Telecommunication Union, 2025.

[6] Armstrong, C., Thresh, L., Murphy, D., and Kearney, G., "A perceptual evaluation of individual and non-individual HRTFs: A case study of the SADIE II database," *Applied Sciences*, 8(11), p. 2029, 2018.

[7] Letowski, T., "Sound quality assessment: con-

cepts and criteria," in *Audio Engineering Society Convention 87*, Audio Engineering Society, 1989.

[8] Reardon, G., Roginska, A., Flanagan, P., Calle, J. S., Genovese, A., Zalles, G., Olko, M., and Jerez, C., "Evaluation of binaural renderers: a methodology," in *Audio Engineering Society Convention 143*, Audio Engineering Society, 2017.

[9] Lee, H., Sochaczewska, K., Wollage, N., Boynton, K., Caruso, A., Lindberg, M., Ratcliffe, S., and Willsher, M., "ECHO Project – Exploring the Cinematic Hemisphere for Orchestra," 2025, doi: 10.5281/zenodo.15268471.

[10] Thornberg, R., Charmaz, K., et al., "Grounded theory and theoretical coding," *The SAGE handbook of qualitative data analysis*, 5(2014), pp. 153–69, 2014.

[11] Zacharov, N., *Sensory evaluation of sound*, CRC Press, 2018.

[12] Catic, J., Santurette, S., and Dau, T., "The role of reverberation-related binaural cues in the externalization of speech," *The Journal of the Acoustical Society of America*, 138(2), pp. 1154–1167, 2015.

[13] Best, V., Baumgartner, R., Lavandier, M., Majdak, P., and Kopčo, N., "Sound externalization: A review of recent research," *Trends in hearing*, 24, p. 2331216520948390, 2020.

[14] Rumsey, F., Zieliński, S., Kassier, R., and Bech, S., "On the relative importance of spatial and timbral fidelities in judgments of degraded multichannel audio quality," *The Journal of the Acoustical Society of America*, 118(2), pp. 968–976, 2005.

[15] International Telecommunication Union, "ITU-R BS.2399-0: Methods for selecting and describing attributes and terms, in the preparation of subjective tests," Recommendation BS.2399-0, International Telecommunication Union, Geneva, Switzerland, 2017.

[16] Friedrich, P. and Schönwiesner, M., "Adaptation rate and persistence across multiple sets of spectral cues for sound localization," *The Journal of the Acoustical Society of America*, 157(3), pp. 1543–1553, 2025.