



Audio Engineering Society Conference Paper

Presented at the AES 2026 International Conference on
Audio for Virtual and Augmented Reality and Immersive Games
2026 June 30 – July 3, Paris, France

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

Eliciting the Effectiveness of Binaural Renderer Enhancement on the Horizontal Plane

Garrett Treanor^{1,2}, Shaunak Ranade^{1,2}, Jean-Marc Jot¹, Daphna Harel², and Agnieszka Roginska²

¹*Virtuel Works LLC*

²*New York University*

Correspondence should be addressed to Garrett Treanor (gkt230@nyu.edu)

ABSTRACT

Binaural rendering is central to spatial audio reproduction via headphones and wearable devices, yet systematic evaluation of enhancement techniques remains methodologically inconsistent across the literature. This paper presents and applies a subjective evaluation methodology designed to consistently elicit five perceptual attributes of headphone-based spatial audio: externalization, horizontal elevation fidelity, front/back distinction, room tone, and spectral coloration. The methodology combines absolute judgment and relative comparison protocols, taking advantage of their complementary capabilities to capture both the absolute perceptual quality of individual rendering conditions, and the salience of perceived differences between them. It is applied in a controlled experiment comparing conventional HRTF-based binaural rendering against two enhancement variants that superpose a masked spatially diffuse sound field component into the binaural output. Stimuli were spatialized across eleven azimuthal positions in the horizontal plane using a generic HRTF dataset, and presented over closed-back headphones to participants. The results validate the proposed methodology as a tool for revealing perceptually relevant differences among binaural rendering conditions. Relative comparison tests reveal additional performance difference details between rendering methods. Both enhancement methods significantly improve externalization and front/back distinction relative to unenhanced binaural rendering, with the largest gains at lateral azimuths, and without a statistically significant increase in perceived spectral coloration beyond the baseline effect of HRTF filtering. However, the results do not conclusively indicate whether the binaural rendering methods examined here exhibit or mitigate a "spurious elevation" artifact associated with frontal sound presentation.

1 Binaural Enhancement Processing

Binaural processing technology is essential to the success of modern media production and consumption. It is required for delivering convincing listening experiences via head-worn playback devices (some of which feature listener head-pose tracking and compensation).

Media content includes music or movie soundtracks recorded in multichannel immersive audio formats compatible with flexible headphone or loudspeaker listening conditions. In these applications, binaural enhancement processing is not just an optional embellishment: it is necessary to realize faithful reproduction of audio "depth" along front-back and radial dimensions.

Applicative context evolution motivates binaural rendering technology innovation, along with a general experimental methodology for performance assessment. However, as detailed in [1], existing efforts to evaluate binaural reproduction struggle to account for various real-world conditions, including diverse program content, playback devices, and application scenarios. As a result, it remains difficult to decisively evaluate the extent to which novel enhancement processing methods address psychoacoustic challenges commonly encountered in binaural audio reproduction – namely: externalization, horizontal elevation fidelity, front-back distinction, and spectral coloration [2]. Moreover, as head-worn virtual, augmented, mixed, and extended reality devices gain greater relevance in the consumer market, generic or non-individualized head-related transfer functions (HRTFs) become more widely adopted and relevant to binaural research [3, 4].

1.1 Previously reported challenges of binaural reproduction and proposed solutions

One of the most well-researched binaural psychoacoustic phenomena is externalization, where a sound source is perceived to originate from outside of the head [5]. Frequently, externalization is enhanced through the simulation of real-world environmental cues over headphones, including head-pose tracking [3, 6], the use of virtual loudspeakers (binaural virtualization) [7, 8], virtual acoustic spaces [9], and HRTF individualization [10, 11]. However, since most of these approaches introduce spectral coloration artifacts, it remains critical to assess this potential impact on presented content.

Timbral alterations, such as spectral coloration, to source content are consistently seen as undesirable side-effects of the binauralization and externalization processes, and there have been equally as consistent efforts to make these processes as perceptually transparent as possible for more than 20 years [7]. As spectral cues are critical to localization [12], some coloration is largely unavoidable when applying HRTF-based filters, though the perceived impact may depend on the similarity between the applied filter and the listener’s own HRTFs [13].

While the perceptual components of vertical localization are well understood [14, 15], it has proved challenging to resolve unwanted perceived elevation of binaural content whilst maintaining a convincing binaural experience, especially without individualization [4].

Under static binaural synthesis conditions, elevation perception depends greatly on spectral content [16], but the impact of erroneous vertical localization may also be influenced by azimuth¹, interaural time differences derived from assumed HRTF symmetry [15], and the listener’s impression of the acoustic environment [17]. While the pinnae are largely responsible for the spectral cues known to influence elevation perception, their small size practically limits their spectral modifications to frequencies above 3 kHz [14, 18].

A large number of investigations aimed at mitigating front-back distinction issues utilize dynamic binaural synthesis [4, 9]. While especially relevant for XR technologies, this approach does not address conventional headphone listening. Other proposed methods of resolving front-back confusion have focused on spectral alterations, including HRTF individualization [13, 19], synthetic reverberation cues [17], spectral manipulation of program content [20, 21] or applied HRTF filters [22].

To address the headphone reproduction of recordings intended for loudspeaker-based reproduction, binaural post-processing approaches have been proposed and developed over time [7]. As mentioned previously, virtual loudspeaker simulation has been at the core of these enhancements, though the principles of their implementations have differed. Some methods strive to emulate the natural listening experience [23, 24, 25, 26], to mitigate up-mixing errors [27, 28], or to differentiate the virtualization processing of pre-existing diffuse components (such as reverberation or ambience) [23]. While these methods may vary in reported efficacy, their related studies often fail to capture enough information related to their perceptual impact for meaningful comparisons, leading to a state-of-the-art with valuable work that is difficult to parse and navigate.

2 Perceptual Evaluation: Combining Absolute and Relative Judgments

Building on the framework proposed in [1], we employ a subjective evaluation methodology that assesses binaural enhancement across five perceptual attributes consistently identified as central to headphone-based spatial audio quality: externalization, horizontal elevation fidelity, front/back distinction, room tone, and

¹https://www.lesonbinaural.fr/EDIT/PDF/voix_interieure_binaural_natif_daw.PDF

spectral coloration. This attribute set is robust across content types and processing methods, supporting direct comparison and aggregation of results.

The methodology pairs absolute judgment with relative comparison protocols, which capture complementary aspects of the listening experience. Absolute tests rate individual stimuli on perceptual scales anchored to defined reference points (e.g. "inside the head" versus "completely outside the head" for externalization), providing baseline performance metrics. Relative comparison tests, by contrast, measure perceived differences between paired conditions in a single judgment, revealing *how much* listeners differentiate between alternatives.

This distinction matters for two fundamental reasons. First, the two formats engage different cognitive processes. Direct A/B comparison focuses listeners on differential properties between conditions and often surfaces nuanced distinctions that may not emerge in isolated assessment; it controls for contextual anchoring effects that can bias absolute judgments. Recanzone et al. [29] demonstrated this asymmetry empirically in sound localization, finding that relative judgments yielded substantially finer discrimination thresholds than absolute judgments of the same stimuli, which supports the use of both protocols as complementary rather than redundant measures. Second, relative tests more faithfully match real-world listening: engineers compare before-and-after processing, consumers compare products, and users switch between spatial audio renderings. Relative metrics thus encode the perceptual salience of differences as they manifest in authentic comparative scenarios.

The combined structure of this methodology also illuminates cognitive comfort: the ease and confidence with which listeners form perceptual judgments. In absolute tests, response variability and within-listener consistency index judgment confidence: low variance suggests the attribute is perceived as a stable, salient property, while high variance suggests perceptual ambiguity. Relative tests typically yield lower variance on a given dimension, reflecting the cognitive ease of comparative judgment: deciding "more" or "less" between two stimuli requires less perceptual inference than anchoring each stimulus to an abstract scale. Response patterns across both protocols therefore indicate which attributes listeners find intuitive and accessible (low variance in both) versus which are more subjective or listener-dependent (higher variance, particularly in absolute judgments).

Together, these protocols yield a more complete characterization of binaural quality: not only what differences listeners perceive, but how confidently and consistently they perceive them, and how those perceptions shift between isolated and comparative evaluation.

2.1 Experimental Protocol

26 participants (14 male, 12 female), ranging in age from 22 to 60, with the majority in their 20s, took part in this experiment. All participants were associated with New York University's Music Technology program with prior experience in critical listening, and 25 out of 26 reported prior experience with spatial audio. All participants reported normal hearing. Three test protocols were designed to evaluate the perceptual impact of all processing methods across five attributes: externalization, elevation, front/back distinction, room tone and spectral coloration. 13 participants completed the *Absolute Judgment test* and an additional 13 completed the *Relative Comparison test*, with all 26 participants completing the *Spectral Coloration test*.

2.1.1 Source Stimuli

The study utilized two unique audio samples with contrasting acoustic properties:

- **Female speech:** A voice sample captured in an anechoic environment.
- **Rainstick:** Chosen for its steady, broad-band noise quality with a wide frequency range.

Both audio files had a sampling rate of 96 kHz and 24-bit depth resolution. To prevent listener adaptation, each trial used a randomly selected 5-second excerpt from the source material.

2.1.2 Spatial Reproduction Conditions

The experiment tested four different processing methods:

1. **Mono reference:** Direct monophonic signal
2. **Binaural processing:** Conventional HRTF-based binaural rendering [30]
3. **Binaural with Enhancement #1:** Enhanced binaural processing, variant #1 (see below)
4. **Binaural with Enhancement #2:** Enhanced binaural processing, variant #2.

HRTF filter pairs derived from measurements performed on a Neumann KU-100 dummy head were used for spatial rendering, with diffuse-field compensation applied. The dataset was sourced from the BiLi measurement campaign [31]. Eleven azimuthal positions along the horizontal plane were tested: 0° (front), 15° , 30° , 45° , 60° , 90° , 120° , 135° , 150° , 165° , and 180° .

To keep the experiment duration manageable while preserving adequate spatial sampling, we assumed symmetrical perception between left and right hemispheres. Each participant heard stimuli processed at 0° and 180° , with the other nine positions being randomly distributed across left and right hemispheres. The processed stimuli were played back in a quiet booth ($RT_{60} = 0.2s$) using Audio-Technica ATH-M40x closed-back headphones at each participant's preferred comfortable loudness level.

2.1.3 Binaural rendering enhancement methods

Besides standard binaural rendering, two enhancement processing methods were compared. Both apply the diffuse tail processing principle introduced in [7]. Figure 1 illustrates *Enhancement #1* (demonstrated in [32]), where g_0 denotes an attenuation factor ensuring amplitude normalization. In *Enhancement #2*, the diffuse tail processing module is fed by the output signals of the binaural rendering module.

2.1.4 Absolute Judgment Test Protocol

In the Absolute Judgement test, participants rated audio samples across three perceptual attributes:

1. **Externalization:** "Where do you hear the sound? Rate the degree of externalization relative to the edge of the head" on a five-point scale: "Inside the head," "At the edge of the head," "Slightly outside the head," "Completely outside the head," and "Can't say."

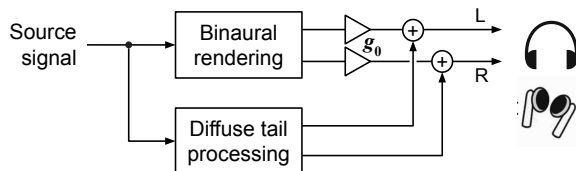


Fig. 1: Binaural rendering enhancement method #1

2. **Horizontal Elevation Fidelity:** "Rate the degree of elevation relative to a horizontal direction of sound arrival" on a five-point scale: "Much higher than a horizontal direction," "Slightly higher than a horizontal direction," "A direction on the horizontal plane," "Slightly lower than a horizontal direction," and "Much lower than a horizontal direction."
3. **Front/Back Distinction:** "Rate the degree of front/back distinction relative to a lateral direction of sound arrival" on a five-point scale: "Distinctly from the front hemisphere," "Probably from the front hemisphere," "Not distinctly from the front or back hemisphere," "Probably from the back hemisphere," and "Distinctly from the back hemisphere."

Each listener was randomly assigned either the female speech or the rainstick sound. For reliability purposes, all conditions were presented in duplicate with randomized trial order, resulting in 68 total trials per participant: 11 spatial positions $\times$ 3 binaural processing methods + 1 mono condition, each repeated twice.

2.1.5 Relative Comparison Test Protocol

The Relative Comparison test presented participants with paired stimuli, asking them to compare stimulus B against stimulus A on four perceptual attributes:

1. **Externalization:** "Rate the degree of externalization of B compared to A" with available responses: "B is more externalized than A," "B is less externalized than A," "B and A have the same degree of externalization," or "I can't tell."
2. **Horizontal Elevation Fidelity:** "Rate the degree of elevation of B compared to A" with available responses: "B has a higher elevation angle than A," "B has a lower elevation angle than A," "B and A have the same elevation angle," or "I can't tell."
3. **Front/Back Distinction:** "Rate the degree of front/back distinction of B compared to A" with available responses: "B is more distinctly localized in the front/back hemisphere than A," "B is less distinctly localized in the front/back hemisphere than A," "Neither A nor B is more distinctly localized in a front/back hemisphere," or "I can't tell."
4. **Room Tone:** "Rate the presence of room tone in B compared to A" with available responses: "B has a more pronounced room tone than A," "B has a less pronounced room tone than A," "Neither A nor B has a more pronounced room tone," or "I can't tell."

Six comparison pairs were tested: 1) Mono vs Binaural; 2) Mono vs Enhancement #1; 3) Mono vs Enhancement #2; 4) Binaural vs Enhancement #1; 5) Binaural vs Enhancement #2; 6) Enhancement #1 vs Enhancement #2. To counterbalance order effects, each pair appeared in both AB and BA configurations (also serving as reliability checks), resulting in 134 trials total: 12 comparison pairs per spatial position (6 pairs, 2 orders) across 11 locations, plus 2 mono-vs-mono control trials.

2.1.6 Spectral Coloration Test Protocol

The Spectral Coloration test examined timbral differences through pairwise comparisons between processed and reference audio, with the unprocessed mono signal always serving as the reference. The available responses were: "Significant difference in spectral coloration," "Slight difference in spectral coloration," or "No difference in spectral coloration." Three pairs were assessed: 1) Mono vs Binaural; 2) Mono vs Enhancement #1; 3) Mono vs Enhancement #2. Following the same counterbalancing approach as in the Relative Comparison test, pairs appeared in both AB and BA order to mitigate ordering bias, producing 68 trials in total: 6 comparisons per spatial location (3 pairs, 2 orders) across 11 positions, plus 2 mono-vs-mono control trials.

2.1.7 Experimental Controls

The trial order was randomized within each test to reduce systematic biases and learning effects. To evaluate response consistency and enhance reliability, all experimental conditions appeared twice across the tests. For the Absolute Judgment protocol, this repetition was implemented via duplicated stimuli. For the Relative and Spectral Coloration protocols, the AB/BA counterbalancing inherently provided the duplicate presentations. Participants received either the female speech or the rainstick stimulus (assigned randomly) to examine whether results generalized across different audio content. Sessions lasted between 1.5 and 2 hours per participant, with variation depending on test assignment and individual pacing.

3 Listening Assessment Results and Statistical Analysis

To examine differences in perceived attributes among processing methods, Generalized Additive Models (GAMs) were employed for both the Absolute and Relative test analyses [33]. This section presents findings

related to externalization, horizontal elevation fidelity, front/back distinction, spectral coloration, and room tone, under both the Absolute and Relative testing conditions outlined in Section 2.

In the Absolute test, GAMs were fitted to attribute rating responses (1–5) for each processing method individually as a function of azimuth. For the Relative test, listener responses were analyzed using ordinal GAMs implemented in the `mgcv` package in R. For each perceptual outcome, the model included the comparison pair as a fixed effect, penalized smooth functions of azimuth to capture spatial trends, and a random effect for listeners to account for individual differences. To quantify difference between methods, we computed a perceived difference metric defined by $P(\text{second method effect}) - P(\text{first method effect})$, where positive values indicate that the second method was more enhanced for certain attributes than the first. Uncertainty was estimated by simulating 1,000 draws from the model's coefficient covariance matrix, from which predicted probabilities and perceived difference scores were derived, with 95% confidence intervals obtained from the empirical quantiles of the simulated distribution.

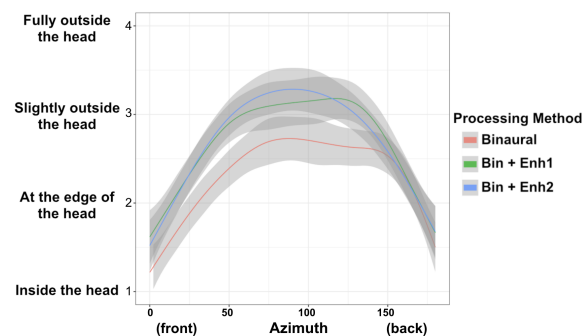


Fig. 2: Absolute externalization judgments vs. azimuth in the horizontal plane (95% CI)

3.0.1 Externalization Results

As shown in Fig. 2, all three processing methods follow a similar parabolic trend, with externalization ratings rising from 0° azimuth, peaking around 90° – 120°, and falling toward 180°. At 0° azimuth, all three methods converge near the “inside the head” boundary, with a slight enhancement advantage. As azimuth increases, Enhancement #1 (*Bin + Enh1*) and Enhancement #2 (*Bin + Enh2*) both diverge notably from Binaural, reaching ratings exceeding “slightly outside the

head,” while ratings for Binaural remain consistently lower. Beyond 165° , all three methods converge again at lower externalization ratings with narrowing confidence bands, suggesting that all processing methods elicit similarly poor externalization in the rear.

In contrast, the relative test results for externalization indicate externalization enhancement in both front and back positions (Fig. 3). A positive perceived difference value indicates that the second processing method in a given pair was perceived as more externalized than the first method, while a value of zero indicates no perceived externalization difference between the two. Any values below zero indicate that the first processing method was perceived as more externalized than the second. As shown, all three comparison pairs produced perceived externalization curves well above zero across all azimuths, indicating that in each case the second method was more externalized. Both enhancement methods elicited significantly more externalization than Binaural across all azimuths (*Bin + Enh1 vs Bin*: $\Delta_{ext} = 0.51$, 95% CI [0.39, 0.62], $p < 0.001$; *Bin + Enh2 vs Bin*: $\Delta_{ext} = 0.55$, 95% CI [0.43, 0.67], $p < 0.001$). This difference was present across the full azimuth range but was most pronounced at lateral angles, peaking around 45° and 150° . At 0° azimuth, both enhancement methods were not as effective, suggesting that externalization in the front remains challenging regardless of the processing method. When comparing the two enhancement methods directly, there was a statistically significant difference in perceived externalization, with Enhancement #2 being more externalized than Enhancement #1 ($\Delta_{ext} = 0.22$, 95% CI [0.12, 0.33], $p < 0.001$).

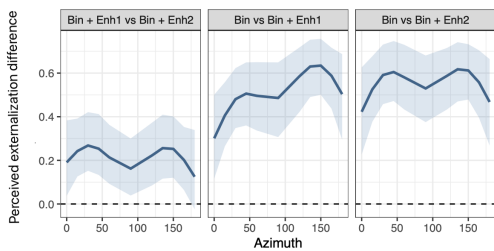


Fig. 3: Perceived externalization difference (95% CI)

Figure 4 indicates the predicted probability of each response category aggregated across all presented positions for each comparison pair, calculated by the proportionality of responses. Across all three pairs, the

probability of the first method being more externalized than the second method was negligible. When comparing Enhancement #1 to Enhancement #2, about 25% of responses indicated that Enhancement #2 was more externalized than Enhancement #1, and 75% of responses indicated that they were equally externalized. This suggests that while Enhancement #2 was statistically more externalized, the perceptual difference between the two enhancement methods was modest for many listeners. In contrast, for both *Bin vs Bin + Enh1* and *Bin vs Bin + Enh2*, the "equal" and "second method preferred" responses were more evenly split with a consistent preference for the enhancements, indicating a more consistent increase in externalization for the enhancement methods compared to unenhanced binaural rendering.

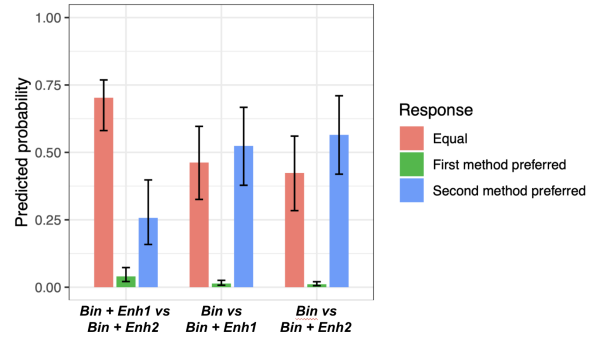


Fig. 4: Externalization results: predicted probability (aggregated across all presented azimuths)

3.0.2 Horizontal Elevation Fidelity Results

As seen in Fig. 5, all three processing methods exhibit similar elevation ratings at 0° azimuth. As azimuth increases, the ratings for the three methods remain close and follow a similar trend. Between $120^\circ - 150^\circ$, however, the Binaural condition exhibits a visible dip in perceived elevation, while Enhancement #1 and Enhancement #2 maintain comparatively stable ratings closer to the horizontal plane. Beyond 165° , all three methods show a slight upward trend in perceived elevation.

The relative results (Fig. 7) revealed a somewhat different pattern from the absolute test. Both enhancement methods were associated with higher perceived elevation relative to Binaural across all azimuths. When compared to Binaural, Enhancement #1 exhibits a gradual increase in elevation ratings from 0° to approximately

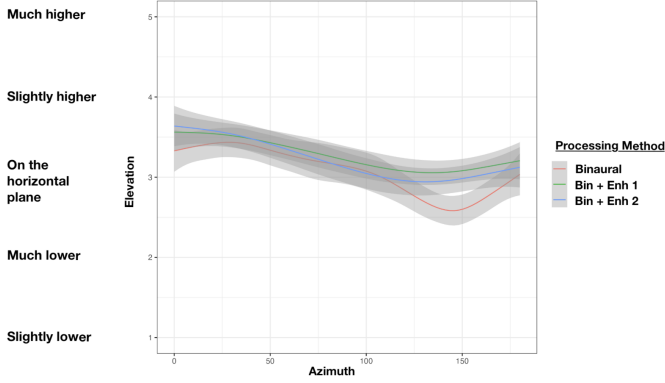


Fig. 5: Absolute elevation judgments (95% CI)

60°, after which the difference levels off. Enhancement #2, by contrast, shows the opposite trend — beginning with a comparatively higher elevation rating at 0° and gradually declining toward 180°. When comparing the two enhancement methods directly, the difference was not statistically significant ($\Delta_{elev} = 0.07$, 95% CI $[-0.03, 0.18]$, $p = 0.18$), with confidence bands crossing zero across the entire azimuth range, suggesting that the two enhancement methods were largely equivalent in terms of horizontal elevation fidelity.

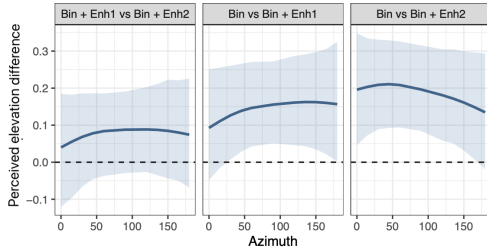


Fig. 7: Perceived elevation difference (95% CI)

3.0.3 Front-Back Distinction Results

Per the absolute judgment results shown in Fig. 6, all three processing methods were largely perceived similarly in terms of front-back distinction until 120° – 150°, where the Binaural condition has a marked departure from the two enhancements. At 0° azimuth, all three methods clustered around the “not distinct” mark. As azimuth increases, all three methods trend upward toward the “probably back” response. Enhancement #2 evoked slightly higher ratings than Enhancement #1

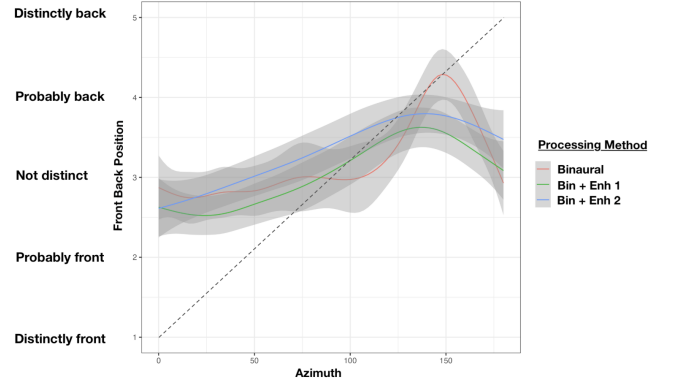


Fig. 6: Absolute front/back distinction judgments (95% CI, expected trend line shown)

across all azimuths. Around 135° – 150°, the Binaural condition exhibits a pronounced peak exceeding “probably back” before declining sharply toward 180°, while Enhancement #1 and Enhancement #2 show comparatively flatter peaks in the same region.

In relative comparison judgments (Fig. 8), both Enhancement #1 ($\Delta_{F/B\ distinction} = 0.21$, 95% CI $[0.11, 0.31]$, $p < 0.001$) and Enhancement #2 ($\Delta_{F/B\ distinction} = 0.18$, 95% CI $[0.09, 0.28]$, $p < 0.001$) were perceived as significantly more distinct than Binaural, with difference curves remaining consistently above zero across both front and back positions. When comparing the two enhancement methods directly, the difference was negligible and not statistically significant ($\Delta_{F/B\ distinction} = 0.04$, 95% CI $[-0.05, 0.13]$, $p = 0.40$), indicating that there is not a difference in perceived front-back distinctiveness between the two enhancement methods.

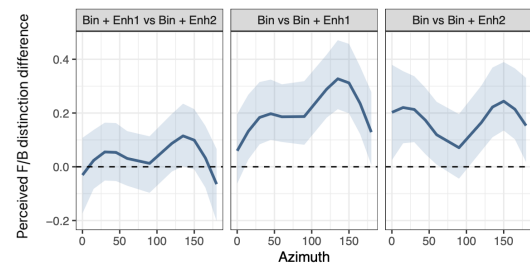


Fig. 8: Perceived degree of front/back distinction difference (95% CI).

3.1 Room Tone Results

As illustrated in Fig. 9, both enhancement methods introduced a statistically significant difference in perceived room tone relative to Binaural, with Enhancement #1 ($\Delta_{room\ tone} = 0.49$, 95% CI [0.37, 0.61], $p < 0.001$) and Enhancement #2 ($\Delta_{room\ tone} = 0.53$, 95% CI [0.42, 0.64], $p < 0.001$) having a greater presence of room tone compared to unenhanced binaural rendering. When comparing the two enhancement methods directly, Enhancement #2 exhibited a higher presence of room tone over Enhancement #1 ($\Delta_{room\ tone} = 0.26$, 95% CI [0.16, 0.37], $p < 0.001$), though this difference was not as noticeable as the difference each enhancement method had over unenhanced binaural rendering.

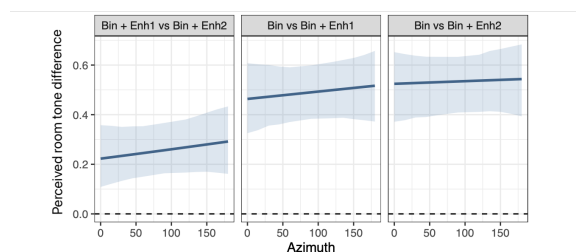


Fig. 9: Perceived room-tone difference (95% CI)

3.1.1 Spectral Coloration Results

As shown in Fig. 10, perceived spectral coloration increased substantially from the Mono condition to Binaural, with both enhancement methods introducing incrementally greater coloration beyond that. Statistical analysis confirmed that all three processing methods differed significantly from Mono in perceived coloration, though there was no statistically significant difference found between Enhancement #1 and Enhancement #2 ($p = 0.06$), or between Enhancement #1 and Binaural ($p = 0.09$). Stimulus type had no significant effect on coloration ratings ($p = 0.654$), indicating these trends were consistent across both stimuli tested.

4 Discussion

In view of the experimental results reported in the previous section, we consider the following question: can the proposed methodology reveal significant and relevant insights about the compared effectiveness of the binaural rendering conditions under study? The results suggest that it can, and on multiple dimensions simultaneously.

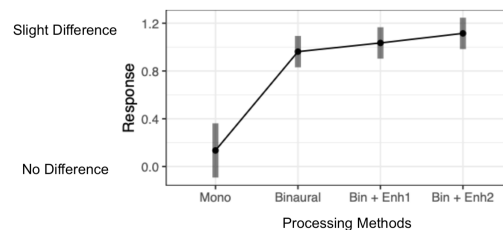


Fig. 10: Perceived coloration difference, aggregated across all presented azimuths (95% CI)

Differences in externalization effectiveness are revealed, as well as their variation according to the incoming direction of the presented sound stimulus. Both enhancement methods deliver a noticeable improvement in externalization for more than half of the listener population. This improvement is most pronounced for azimuths centered around 45 or 135 degrees, and slightly superior for enhancement method 2. Critically, this is achieved without significantly increasing perceived spectral coloration beyond the existing effect of HRTF filtering, relative to the original sound. Practically, this is important for content creators and device manufacturers for whom timbral transparency is an important design constraint. In direct comparison between rendering methods, externalization improvements are accompanied by an increase in reported "room tone." This may reveal a timbral artifact (a type of spectral coloration), or may indicate that the term "room tone" is in fact lexically correlated with the term "externalization." Disentangling these interpretations would require follow up experiments with explicit room tone anchoring and more controlled stimulus design.

In addition to improved externalization, a desirable effect of binaural renderer enhancement is the reduction of localization ambiguity within a "cone of confusion" (as defined, for instance, in [34]). Relative test results reported in section 3 indicate that both enhancement methods indeed improve front/back distinction (confidence). The results also indicate that the proposed experimental methodology can reveal subtle effects in the horizontal preservation performance of various binaural rendering conditions. Despite this, the two presented binaural renderer enhancement methods do not conclusively remedy or mitigate this phenomenon.

The complementary design of the absolute and relative judgment test protocols proved beneficial in practice.

In the absolute judgment test, all three binaural processing methods followed a broadly similar trend across azimuth for externalization. This pattern required careful inspection to distinguish, given the overlapping confidence intervals and several azimuths. The relative comparison test made the same underlying differences more legible. The perceived difference curves consistently above zero across all azimuths made it clear that both enhancement methods outperformed unenhanced binaural rendering. This is consistent with the expectation that comparative judgment reduces the cognitive burden of anchoring responses to abstract (or implicit) scales. For elevation, the two types of tests told a slightly different story. The absolute test showed the three methods were close at all azimuths, while the relative test showed that both enhancement methods resulted in higher perceived elevation than binaural. Together, these observations suggest that absolute tests are good at characterizing the overall perceptual profile of a rendering method and detecting gross differences, while relative tests provide a greater sensitivity to fine distinctions and are more robust to individual listener anchoring biases. Presenting both protocols yields a more complete characterization of binaural rendering performance than either approach alone.

5 Conclusion

This paper proposed and implemented a multi-protocol subjective evaluation methodology for assessing binaural rendering enhancement techniques across five perceptual dimensions: externalization, horizontal elevation fidelity, front/back distinction, room tone, and spectral coloration. The methodology combines absolute judgment and relative comparison tests, leveraging their complementary strengths to characterize both the level and consistency of perceived differences across binaural rendering conditions.

Applying this framework to a controlled experiment comparing conventional HRTF-based binaural rendering against two enhancement methods, we find that both enhancements consistently and significantly improve externalization and front/back distinction relative to unenhanced binaural rendering. These improvements are most pronounced at lateral azimuths (around 45° and 135°), and persist across both stimuli tested (speech and rainstick). Enhancement #2 demonstrated a modest but statistically significant advantage over Enhancement #1 in perceived externalization, while

the two methods were largely indistinguishable in elevation and front/back distinction. Critically, neither enhancement significantly increased perceived spectral coloration beyond the baseline effect of HRTF filtering, though both introduced a measurable increase in perceived room tone. This finding invites further investigation into whether 'room tone' judgments reflect a timbral artifact or a perceptual correlate of externalization itself.

Several directions remain open for future investigation. On the stimulus side, extending the content set beyond anechoic speech and broadband noise (rainstick) to include music, environmental sound, and reverberant material would strengthen the generalizability of conclusions. Additional binaural reproduction technologies, including individualized HRTFs and head-tracking conditions, and a wider range of listening environments warrant evaluation within the same framework. The current perceptual attribute set could be expanded to include lateralization, apparent source width, and sharpness or focus, each of which carries practical relevance for immersive audio applications. Finally, revisiting the wording of survey questions to see whether terms such as "plausibly" or "acceptably" better anchor listener judgment may sharpen the sensitivity and interpretability of the methodology in future experiments.

References

- [1] Jot, J.-M., Treanor, G., Ranade, S., Li, Z., Harel, D., and Roginska, A., "A subjective evaluation methodology for binaural enhancement," in *Audio Eng. Soc. Conf. on Headphone Technology*, 2025.
- [2] Reardon, G., Calle, J. S., Genovese, A., Zalles, G., Olko, M., Jerez, C., Flanagan, P., and Roginska, A., "Evaluation of Binaural Renderers: A Methodology," in *Audio Eng. Soc. 143rd Conv.*, 359, 2017.
- [3] Hendrickx, E., Stitt, P., Messonnier, J.-C., Lyzwa, J.-M., Katz, B. F., and de Boishéraud, C., "Influence of head tracking on the externalization of speech stimuli for non-individualized binaural synthesis," *J. Acoust. Soc. Am.*, 141(3), pp. 2011–2023, 2017.
- [4] Frankenbach, C., Herzog, C., Seitz, J., and Fels, J., "Evaluating Localization Accuracy of Dynamic Binaural Synthesis: Differences Between Individual, Estimated, and Generic HRTFs," *J. Audio Eng. Soc.*, 73(1/2), pp. 5–23, 2025.
- [5] Best, V., Baumgartner, R., Lavandier, M., Majdak, P., and Kopčo, N., "Sound Externalization: A Review of Recent Research," *Trends in Hearing*, 24, pp. 1–14, 2020.

- [6] Brimijoin, W. O., Boyd, A. W., and Akeroyd, M. A., "The Contribution of Head Movement to the Externalization and Internalization of Sounds," *PLoS ONE*, 8(12), pp. 1–12, 2013, doi:10.1371/journal.pone.0083068.
- [7] Jot, J.-M., Lukin, A., and Landschoot, C., "Binaural Externalization Processing - from Stereo to Object-Based Audio," in *Audio Eng. Soc. 153rd Conv.*, 2022.
- [8] Merimaa, J., "Modification of HRTF Filters to Reduce Timbral Effects in Binaural Synthesis," in *Audio Eng. Soc. 127th Conv.*, 7912, 2009.
- [9] 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," *J. Audio Eng. Soc.*, 49(10), pp. 904–916, 2001.
- [10] Hartmann, W. and Wittenberg, A., "On the externalization of sound images," *J. Acoust. Soc. Am.*, 99, pp. 3678–88, 1996, doi:10.1121/1.414965.
- [11] Kim, S.-M. and Choi, W., "On the externalization of virtual sound images in headphone reproduction: A Wiener filter approach," *J. Acoust. Soc. Am.*, 117(6), pp. 3657–3665, 2005, doi:10.1121/1.1921548.
- [12] Blauert, J., *Spatial Hearing: The Psychophysics of Human Sound Localization*, 1996.
- [13] Jenny, C. and Reuter, C., "Usability of Individualized Head-Related Transfer Functions in Virtual Reality: Empirical Study With Perceptual Attributes in Sagittal Plane Sound Localization," *JMIR Serious Games*, 8, 2020.
- [14] Macpherson, E. A. and Sabin, A. T., "Vertical-plane sound localization with distorted spectral cues," *Hearing Research*, 306, pp. 76–92, 2013, doi:10.1016/j.heares.2013.09.007.
- [15] Genovese, A., Juras, J., Miller, C., and Roginska, A., "The Effect of Elevation on ITD Symmetry," in *Audio Eng. Soc. Conf. on Headphone Technology*, 2016.
- [16] Hofman, P. M., Riswick, J. G. V., and Opstal, A. J. V., "Relearning sound localization with new ears," *Nature neuroscience*, 1(5), pp. 417–421, 1998.
- [17] Carlini, A., Bordeau, C., and Ambard, M., "Auditory localization: a comprehensive practical review," *Frontiers in Psychology*, 15, 2024, doi:10.3389/fpsyg.2024.1408073.
- [18] Avendano, C., Algazi, V. R., and Duda, R. O., "A head-and-torso model for low-frequency binaural elevation effects," in *Proc. Workshop Appl. Sig. Proc. Audio and Acoustics (WASPAA)*, pp. 179–182, 1999, doi:10.1109/ASPAA.1999.810879.
- [19] Oberem, J., Richter, J.-G., Setzer, D., Seibold, J., Koch, I., and Fels, J., "Experiments on Localization Accuracy with Non-Individual and Individual HRTFs Comparing Static and Dynamic Reproduction Methods," in *Fortschritte der Akustik: DAGA 2018*, 2018.
- [20] Brimijoin, W. O. and Akeroyd, M., "The Effects of Hearing Impairment, Age, and Hearing Aids on the Use of Self-Motion for Determining Front/Back Location," *J. Am. Acad. Audiol.*, 27, 2016, doi:10.3766/jaaa.15101.
- [21] Zhang, P. X. and Hartmann, W. M., "On the ability of human listeners to distinguish between front and back," *Hearing Research*, 260(1), pp. 30–46, 2010, doi:https://doi.org/10.1016/j.heares.2009.11.001.
- [22] Frank, M. and Zotter, F., "Simple Reduction of Front-Back Confusion in Static Binaural Rendering," in *Fortschritte der Akustik - DAGA*, pp. 329–332, 2018.
- [23] Jot, J.-M. and Avendano, C., "Spatial Enhancement of Audio Recordings," in *Audio Eng. Soc. 23rd Intl. Conf.*, 2003.
- [24] Davidson, G., Darcy, D., Fielder, L., Shuang, Z., Graff, R., Breebaart, J., and Crum, P., "Design and Subjective Evaluation of a Perceptually-Optimized Headphone Virtualizer," in *Audio Eng. Soc. 140th Conv.*, 2016.
- [25] Jot, J.-M. and Lee, K. S., "Augmented Reality Headphone Environment Rendering," in *Audio Eng. Soc. Int. Conf. Audio for Virtual and Augmented Reality*, 2016.
- [26] Werner, S., Klein, F., Mayenfels, T., and Brandenburg, K., "A Summary on Acoustic Room Divergence and its Effect on Externalization of Auditory Events," in *Proc. IEEE 8th Int. Conf. Quality Mult. Exp. (QoMEX)*, 2016.
- [27] Breebaart, J. and Schuijers, E., "Phantom Materialization: A Novel Method to Enhance Stereo Audio Reproduction on Headphones," *IEEE Transactions on Audio, Speech, and Language Processing*, 16(8), pp. 1503–1511, 2008, doi:10.1109/TASL.2008.2002983.
- [28] Goodwin, M. and Jot, J.-M., "Binaural 3-D audio rendering based on spatial audio scene coding," in *Audio Eng. Soc. 123rd Conv.*, 2007.
- [29] Recanzone, G. H., Makhamra, S. D. D. R., and Guard, D. C., "Comparison of relative and absolute sound localization ability in humans," *J. Acoust. Soc. Am.*, 103(2), pp. 1085–1097, 1998, doi:10.1121/1.421222.
- [30] Hammershøi, D. and Møller, H., "Methods for binaural recording and reproduction," *Acustica United with Acta Acustica*, 88(3), pp. 303–311, 2002.
- [31] Carpentier, T., Bahu, H., Noisternig, M., and Warusfel, O., "Measurement of a Head-Related Transfer Function Database with High Spatial Resolution," in *7th EAA Forum Acusticum*, Kraków, Poland, 2014.
- [32] Landschoot, C. and Jot, J.-M., "Colorless Binaural Externalization Processing Demonstration," <https://www.chrislandschoot.com/binaural-externalization>, 2025.
- [33] Hastie, T. J., *Generalized Additive Models*, Routledge, New York, 2017, ISBN 978-0-203-75378-1, doi:10.1201/9780203753781.
- [34] Blauert, J. and Lindemann, W., "Auditory spaciousness: Some further psychoacoustic analyses," *J. Acoust. Soc. Am.*, 80(2), pp. 533–542, 1986, doi:10.1121/1.394048.