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, Sorbonne University/d'Alembert & IRCAM/STMS, 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.

Audiovisual Coherence Thresholds for Direct-to-Reverberant Energy Ratio

Peter Lindahl[1], Johannes M. Arend[1], and Nils Meyer-Kahlen[1, 2]

[1]*Acoustics Lab, Department of Information and Communications Engineering, Aalto University, Finland*
[2]*Audio Experience Design Group, Dyson School of Design Engineering, Imperial College London, United Kingdom*

Correspondence should be addressed to Nils Meyer-Kahlen (nils.meyer-kahlen@aalto.fi)

ABSTRACT

Holographic calling systems aim to create the perceptual illusion that a remote interlocutor is physically present by combining augmented reality (AR) visualizations with spatial audio rendering. A key challenge in such systems is achieving audiovisual coherence when room-acoustic properties must be inferred, since this can lead to greater inaccuracies than when they are estimated from measurements. In this study, we investigate perceptual tolerance to mismatches in the direct-to-reverberant energy ratio (DRR), a critical cue for auditory distance perception. We conducted a listening experiment in which participants judged whether the audio and the visual presentation of a stimulus were coherent using a yes/no task. Stimuli included both a volumetric capture of a speaking human avatar, rendered at the correct direct sound level, and a loudspeaker reproducing wideband noise. For the loudspeaker, level roving was introduced to assess the influence of intensity cues on listener decisions. Results show that audiovisual coherence is maintained for half of the presentations within a DRR mismatch range of approximately 3 dB for too dry and 4.3 dB for too reverberant renderings. Within the limited number of participants included in the study, no evidence for significant differences was found between speaking avatars and loudspeakers reproducing noise, nor between different ranges of level roving in the loudspeaker condition. Nevertheless, the findings help to understand the consequences of DRR estimation mismatches for holographic AR calling experiences.

1 Introduction

The aim of a holographic calling system is to make a remote interlocutor appear to be physically present in a user's space. This involves presenting a three-dimensional visual rendering of the remote person using augmented reality (AR) glasses, with either semi-transparent displays or camera-based pass-through technology. Such visual rendering should be accompanied by position-dynamic binaural audio rendering. To create the correct auditory impression, this rendering should also be *room-adaptive*, i.e., it should include room acoustics that match the features of the user's current environment [1].

The idea of room adaptation naturally raises the question of how accurately room acoustic features need to be matched. This is particularly relevant given that practical room-adapted rendering will likely not be able to rely on dedicated acoustic measurements of a space; instead, the room acoustics must be inferred blindly, for example, based on active sound sources present in the room [2], visual input [3], or both [4].

It has been shown that the perceptual requirements for room adaptation depend on the precise perceptual aim, which, in turn, determines the appropriate test paradigm. Several paradigms for evaluating AR audio rendering have been proposed in recent years. The strictest of these is authenticity, which aims for a virtual rendering that is perceptually indistinguishable from its real counterpart [5]; a goal that is rarely achievable in practice. An alternative goal is to create so-called auditory reality illusions (or just "auditory illusions" in short). An auditory illusion occurs when a virtual sound source is believed to be real [6]; possible test paradigms include plausibility [7] or transfer-plausibility [6]. It has been shown that, even with moderate mismatches between real and virtual room acoustics, auditory illusions can be achieved with a certain probability, even in cases where authenticity cannot [6]. However, one may argue that creating such illusions is still too high a bar for evaluating practical holographic calling systems: when the visual rendering of a hologram is still far from creating a *visual* reality illusion, why should the auditory rendering be expected to do so? Another, less demanding goal could therefore be to strive for audiovisual coherence between the visual and auditory presentation. This is the approach adopted in the present experiment.

In a recent study, listeners were asked to rate a virtual rendering of a loudspeaker in terms of several attributes, including some specifically related to audiovisual coherence [8]. Subjects were, for example, asked "Do you have the impression the sound is coming from the loudspeaker?" and "Does the sound field integrate well into the room?", which reveals the two distinct aspects of source and room coherence (called audiovisual source and room plausibility in [8]). Ratings were performed on a six-point scale, while the reverberation time was modified. Similarly, in [9], the direct-to-reverberant energy ratio (DRR) was varied when showing a physical loudspeaker and being asked "Do you have the impression that the loudspeaker you see in the front is playing? Even small differences of 1.5 dB were sufficient to elicit differences in ratings.

Here, we are also interested in studying DRR. We believe this to be especially relevant for holographic calling systems, as, on the one hand, blind estimation of DRR is difficult to achieve with high precision, and, on the other hand, it is widely considered a crucial and critical cue for auditory distance perception [10, 11, 12, 13]. Therefore, we hypothesize that, despite the presence of a visual capture effect that draws auditory events to visual presentations even in distance [14, 15], a DRR mismatch above a certain threshold disrupts audiovisual source coherence; we aim to determine the threshold at which this occurs. To this end, we conducted an experiment in which the DRR of dynamic, individualized binaural renderings is manipulated. As opposed to [9], we not only vary DRR while showing a loudspeaker, but also present a rendering of a human avatar based on a volumetric point-cloud video, as it might be seen in a holographic calling system. Moreover, we employ a yes/no task rather than a rating task—a more traditional audiovisual coherence paradigm, typically used in experiments investigating mismatches in direction [16, 17].

Apart from DRR, sound intensity is often considered an important distance cue, specifically for relative distance judgments [12]. When changing the DRR by increasing or decreasing the levels of the direct or reverberant sound components, the overall level changes as well. Therefore, it is important to consider how the reproduction level might confound the results of our experiment. It has been shown that humans possess a certain internal reference for the true absolute level of speech, especially for familiar speakers [18]. We believe that for the most convincing rendering, holographic calling technologies should use calibrated chains, such that the direct sound level of speech is equivalent to what would be heard in the real world. Consequently, we do the same in the avatar trials and choose to present the avatar speech conditions at a constant, calibrated direct-sound level corresponding to that of the real human speaker at the listener's position. This choice implies that intensity cues are available to the extent that listeners possess an internal reference of the absolute speech level.

In the loudspeaker conditions, noise bursts were rendered. There can be no absolute internal level reference for such a stimulus; a loudspeaker might play noise bursts at (almost) any level. However, participants may adapt to expect a certain level over the course of the experiment and thereby learn to use intensity cues in their judgments of audiovisual coherence. To test whether this is an issue, we included two loudspeaker conditions: In the first one, the source level was roved within ± 3 dB, mimicking the natural level variability found in the used speech trials. In the second condition, the roving range was increased to ± 8 dB. If listeners formed



(a) Alignment procedure. **(b)** BRIR measurement.

Fig. 1: Screenshots from within the HMD during alignment and individual BRIR measurement.

an internal level reference throughout the experiment, such increased roving would be expected to widen the range of DRR offsets that lead to a coherent audiovisual percept.

In summary, our goal is to answer three research questions: (RQ1) What range of DRR mismatches leads to a breakdown of audiovisual coherence in an AR holographic rendering of a human speaker? (RQ2) Are the limits different when showing a loudspeaker that plays noise bursts? (RQ3) Does the amount of roving introduced to the noise signal change the coherence limits?

2 Methods

2.1 Setup

The listening experiment was conducted in the ITU-R BS.1116-1 compliant listening room "Otala" at the Aalto Acoustics Lab. The DRR and reverberation time (RT) values of the room are shown in Tab. 1. They were obtained by analyzing a binaural room impulse response (BRIR) made at a source-listener distance of 1.5 m, at the same position where source and listener are located during the experiment. The reported RT values are T_{20} estimates.

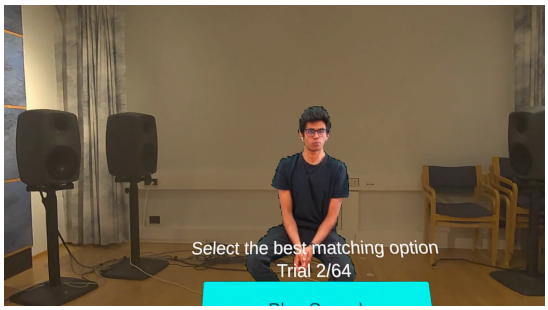
The room was seen through the color cameras and displays of a Meta Quest 3 headset operating in pass-through mode. At the start of the experiment, participants calibrated the virtual environment by positioning a virtual sphere at the center of the coaxial Genelec 8331A loudspeaker present in the room. The prompt shown to participants is shown in Fig. 1a. This location also served as the reference point for the holograms' mouth.

In the next step, BRIR measurements were conducted by each individual participant. We decided to use individual BRIRs, as DRR differences can occur even between different people, due to anatomical differences. Previously, we had assessed the magnitude of such inter-individual differences by computing the directivity index (DI) for a frontal source for each of 60 HRTFs [19]. The DI corresponds to the DRR that would be measured if the room response constituted a diffuse field. Its standard deviation was 0.92 dB, the largest difference to the mean DI was 2.1 dB.

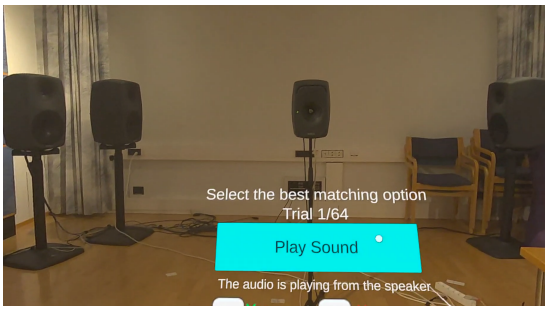
The measurements used MEMS microphones placed in blocked ear canals using the silicone version [20] of the PIRATE in-ear microphones [21], covering source azimuths from $-35°$ to $35°$. Logarithmic sweeps were played through the loudspeaker, with participants turning their heads in $5°$ increments after each measurement, according to instructions displayed in the virtual scene shown in Fig. 1b. The measurements were initiated and controlled from a Matlab script; all real-time

Table 1: RT and DRR in octave bands of the test room without any DRR variations.

Freq. (Hz)	RT (s)	DRR (dB)
250	0.216	-13.20
500	0.243	-3.47
1000	0.232	-0.16
2000	0.265	2.73
4000	0.256	4.75
8000	0.229	4.28



(a) Speech condition with avatar shown.



(b) Noise condition with loudspeaker shown.

Fig. 2: Participants' views during listening test trials. (a) Speech condition. (b) Noise condition.

signal processing was done in Pure Data. The head orientation of the participants was tracked using the Quest 3 headset that sent its orientation via Unity and Open Sound Control (OSC) as a User Datagram Protocol (UDP) message to Pure Data. If participants turned their heads more than 2° away from the target angle during measurements, the measurement was repeated. If they turned their head outside of the ±35° measured range during the listening experiment, all playback was faded out. It was assured that such muting occurred before rendering, so that the room decay was still audible.

A pair of MushRoom headphones [22] was used for the listening tests. The headphone transfer function was measured using the same in-ear microphones prior to the test, and an individual minimum-phase equalization filter was applied to compensate for the microphone and the headphone characteristics. After measuring BRIRs and creating the headphone equalization filter, the entire reproduction chain was measured once in the forward-facing orientation, and the level was subsequently adjusted to match that of the BRIR measurement. This ensured that the level of the virtual rendering corresponded exactly to that of the real loudspeaker.

During the experiment, audio was rendered in real time using the *mtx_convolver~* object from the *iemmatrix* library in Pure Data [23], which convolved either noise or speech samples with the measured BRIRs corresponding to the participant's current head orientation. Half of the trials were presented in a reference condition corresponding to the measured room response (0 dB offset), while the remaining trials were presented with a manipulated DRR. DRR manipulation was performed by modifying only the reverberant component of the signal, defined as all energy arriving later than 3 ms after the direct sound. The direct sound of the BRIR remained unmodified. For clarity, we therefore report the results in terms of "reverberation level offset" instead of DRR. The applied offsets ranged from −6 to +10 dB, in 0.5 dB steps. Unmodified and all modified BRIRs were presented in randomized order. The real loudspeaker was never played during the test, of which the participants were unaware.

2.2 Procedure and Conditions

After giving their informed consent to participate and completing the individual alignment and measurement procedure described above, three experimental conditions were tested: one with a speaking avatar and two showing a loudspeaker that produced noise stimuli. They are described in detail in the following paragraphs. In all conditions, participants saw the pass-through AR view and were instructed to judge whether the sound originated from the avatar or loudspeaker in a yes/no task. In each condition, a total of 64 trials was included, resulting in a listening test duration ranging from 20 to 40 minutes, depending on the participant. To control for potential procedural learning effects, half of the participants completed the speech condition first, while the other half began with the noise conditions. At the end of the experiment, they were asked to rate overall difficulty, list used cues, and estimate how many of the loudspeaker trials used the real loudspeaker (which was never active during the test, as explained above).

2.2.1 Speech with Natural Level Variation

To approximate a realistic AR holographic calling scenario, five different holographic speakers were presented with their naturally occurring level differences

preserved. Figure 2a shows the participants' view in the virtual environment during a speech trial. The loudspeaker used for the measurements and the loudspeaker conditions was physically removed from the participants' view. Instead, a holographic avatar representation of the person speaking was shown, rendered using the Depthkit video plugin in Unity [24]. The holograms were captured at a distance of 1.5 m in an anechoic chamber at the Aalto Acoustics Lab [16]; speech was captured with a low-noise 40HF GRAS microphone at the same distance. In the listening experiment, the holograms were rendered at 1.5 m distance as well, using the same visual scale as in the recordings, and reproduced with the recorded speech levels ranging from 59 to 65 dB A-weighted equivalent continuous sound pressure level across the speakers.

To achieve calibrated playback, the loudspeaker output was measured in the anechoic chamber using the exact same hardware and settings as during the experiment. The measurement was conducted using the same microphone as in the recordings, again, placed 1.5 m from the loudspeaker, corresponding to the original recording distance. The level differences between the recorded speech samples and the playback system were then compensated in Pure Data to match the original speech levels. This yielded a rendering with the same on-axis direct sound level that would be observed if the person had spoken in the room.

2.2.2 Noise with Natural Level Roving

To contrast the more ecologically valid hologram condition with a more laboratory-like setup, the loudspeaker conditions were included, see Fig. 2b. In this condition, the physical loudspeaker remained visible and reproduced wideband noise bursts lasting 300 ms, with 10 ms onset and offset ramps.

As the loudspeaker could plausibly operate at any output level, we cannot expect participants to have an absolute internal level expectations as in the case of the speech. However, they may be able to form level expectations throughout the experiment, especially as the correct condition without offset was presented half of the time. To best compare the speech and the loudspeaker conditions, the noise level was roved uniformly within a ± 3 dB range, approximating the variability observed between the different speech samples.

2.2.3 Noise with Large Level Roving

In addition, we presented a condition using noise stimuli with a larger level roving of ± 8 dB. The purpose of this condition was to reduce the possibility of forming level expectations about the loudspeaker/noise stimulus across the experiment. If this effect played a role, we expect to see wider coherence ranges with this stronger roving.

2.3 Participants

A total of 12 participants with a mean age of 26 years (SD = 1.44 years) completed the test (5 female, 7 male). None of them reported any hearing loss, but two reported having tinnitus, one of them only very slightly. Five of the participants were familiar with the concept of DRR. Eight of the participants had academic training in acoustics.

2.4 Data Analysis

2.4.1 Psychometric Function Fit

Psychometric functions were fitted to the binary responses using the Psignifit library in Python [25]. Fits were conducted separately for each participant and at the group level. The probability of responding "not coherent" was modeled as a function of the applied reverb level offset. The analysis was conducted using the default *experiment_type = "yes/no"* settings [25].

Separate psychometric functions were fitted for negative and positive reverberation-level offsets to account for possible asymmetries. Since for positive offsets, the probability of a "not coherent" response increases by increasing the offset, whereas for negative offset, the relationship is reversed, a cumulative Gaussian sigmoid was used for the positive range, and a decreasing cumulative Gaussian variant was used for the negative range.

The increasing psychometric function definition used for the "yes/no" model was

$$\Psi(x;\alpha,\beta,\lambda,\gamma) = \gamma + (1-\lambda-\gamma)\Phi\left(C\frac{x-\alpha}{\beta}\right), \quad (1)$$

$$C = \Phi^{-1}(0.95) - \Phi^{-1}(0.05), \quad (2)$$

where x denotes the stimulus level, α is the threshold parameter (50 % point) and β represents width parameter defined as the distance between the 5% and 95%

response levels. The functions Φ and Φ^{-1} denote the cumulative standard normal distribution and its inverse. Parameters λ and γ denote the upper and lower asymptotes of the psychometric function, respectively.

In addition to fitting the functions to the reverb level offsets, a decreasing Gaussian sigmoid was fitted to responses across the roving level of all noise stimuli to characterize the effect of source level. The probability of responding "not coherent" was modeled as a function of the applied roving gain in dB. This helps assess whether the level played any role.

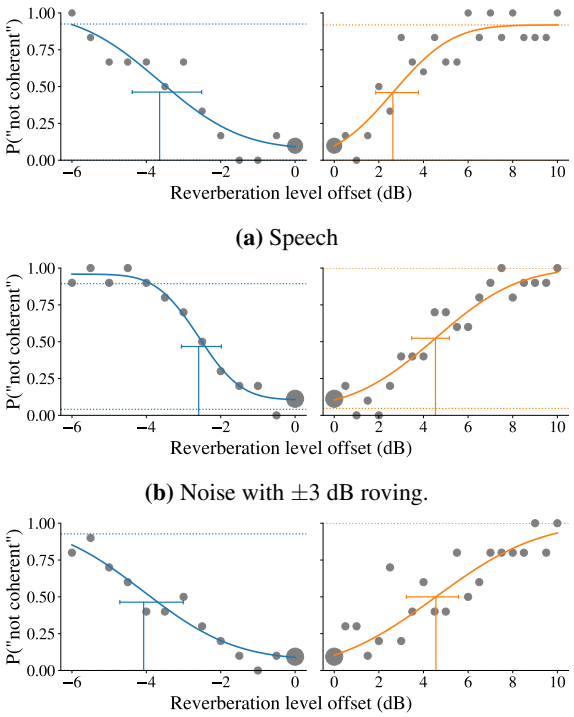
2.4.2 Exclusion

The results of some participants were excluded from each experimental condition because their responses suggested random choice. Participants to be excluded were identified separately for each condition by performing a binomial test on responses in the reference condition with a 0 dB reverb-level offset. Only participants whose response rates in the reference condition were significantly below the 50 % guessing rate were included. This exclusion criterion is illustrated in Fig. 6 in Appendix A, where the individual psychometric function fits are provided for each participant. For example, participant 1 gave approximately equal number of "yes" and "no" responses at 0 dB offset in all conditions. This response pattern suggests that their choices were not reliably related to the stimulus difference, and the participant was therefore excluded.

Based on this criterion, participants 1 and 5 were excluded in the $\pm$8 dB roved noise condition, participant 1 was excluded in the $\pm$3 dB roved noise condition, and participants 1, 8, 10, and 11 were excluded in the speech condition. Additionally, participant 7 was excluded from the speech condition because it was not possible to estimate a psychometric function with a 50% threshold, as they perceived almost all trials as coherent.

2.4.3 Mixed Model

To assess differences between conditions and directions, a mixed model was fit to predict the 50% points. The model had the structure $|x_{50}| \sim$ Condition $\times$ Direction + (1 | Participant), where the test condition (speech, noise with 3 dB roving, noise with 8 dB roving) and direction (positive changes, negative changes) were included as fixed effects and the participant was included as a random effect. x_{50} denotes the offset at which the fitted psychometric functions cross 50%. The model was fit in *R* using the *lme4* package. Marginal means were computed using the *emmeans* package.



(a) Speech

(b) Noise with $\pm$3 dB roving.

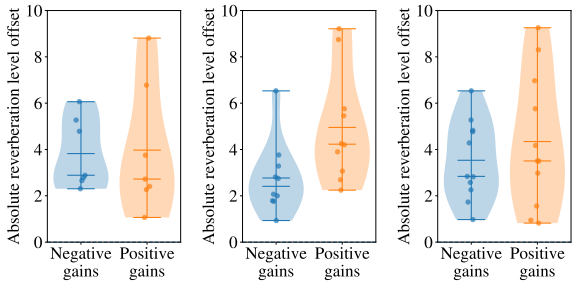
(c) Noise with $\pm$8 dB roving.

Fig. 3: Psychometric functions fit to all participants' data together. Grey dots show proportions of "not coherent" answers per reverberation level offset. The blue curves on the left side show negative reverberation gain offsets, the orange curves on the right show positive offsets. The vertical lines indicate the 50% thresholds and the horizontal lines their 95% confidence intervals.

3 Results

3.1 Detection Thresholds

Figure 3 shows the proportion of responses in which the presentation was perceived as not coherent as a function of reverb level offset for all tested conditions. Overall, the coherence curves are similar across conditions. In all conditions, the reference rendering was accepted in



(a) Speech. **(b)** Noise with ±3 dB roving. **(c)** Noise with ±8 dB roving.

Fig. 4: 50 % points estimated for each participant.

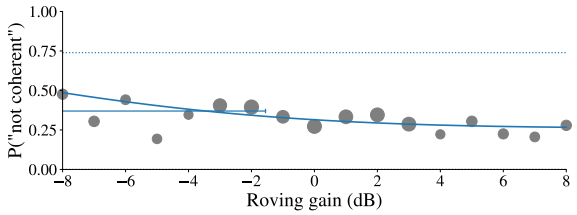


Fig. 5: Psychometric function fit for roving level across noise stimuli responses.

approximately 80% of trials. In the speech condition, a reverb level offset of approximately −4 dB and +3 dB reduced the proportion of responses that were perceived as coherent to 50%, with further decreases observed for larger offsets. The results suggest that errors should be well below 2 dB if no mismatch effect is desired. The noise conditions show a trend toward a higher reverb level offset tolerance for positive offsets compared to speech; however, these differences remain within the 95% confidence intervals, as illustrated in Figs. 3b and 3c.

Figure 4 shows the distributions of 50% points for each individual. The absolute mean 50% points over all conditions were 3 dB for negative and 4.3 dB for positive offsets. The mixed model analysis revealed a significant main effect of direction, $F(1,42.43) = 8.52,\ p = .006$, indicating higher thresholds for positive compared to negative changes. There was no significant main effect of condition, $F(2,43.76) = 0.23,\ p = .792$, and no significant interaction between condition and direction, $F(2,42.43) = 0.53,\ p = .591$. Pairwise comparisons of marginal means showed that the difference between directions was significant in the noise condition with 3 dB roving ($p = .011$), with a difference in marginal means of 1.93 dB, but not in the other two conditions. Nevertheless, a tendency toward higher thresholds for positive changes was found in those parts as well, with differences of 0.89 dB and 1.12 dB for speech and noise with 8 dB roving, respectively.

3.2 Effect of Level Roving

Figure 5 shows the proportion of accepted responses as a function of the applied roving gain. A slight increase in the probability of "not coherent" responses can be observed for lower (quieter) gain values. However, the confidence interval is wide for these values. Thus, results do not show a clear effect of the roving and the playback level on the rates of trials in which participants responded with "coherent".

3.3 Post-Survey Responses

Responses to the survey questions offer some more insight into the subjective experience. When asked how many noise stimuli they thought were playing from the loudspeaker, only one of the 12 participants reported thinking that all of the stimuli were played from the headphones. The other estimates ranged from 10%–80%. Consequently, auditory illusions occurred in some trials. Even if that was not the primary goal, it demonstrates that the rendering did not have obvious defects.

There was no clear consensus regarding which block was perceived as the most difficult. Six participants reported that speech was easier to judge, while four reported it was more difficult than noise. The participants reported focusing on distance and directional cues to make judgments One participant mentioned comparing each stimulus to the one presented before.

4 Discussion

Overall, the thresholds for speech were around 3 dB to 4 dB. However, the results exhibit wide confidence intervals. Some participants found renderings coherent half the time when the DRR mismatch was more than 6 dB. In general, there seems to be a slight asymmetry in that participants tend to respond sooner with "incoherent" at negative offsets (too dry renderings). Mainly, this was found for the loudspeaker noise condition with natural roving (±3 dB). This result is in agreement with the asymmetry of the visual capture effect found in [15], which was stronger for sources rendered to be further away from the participant than the visual anchor.

4.1 Relationship to Just-Noticeable Difference

One could think of the just-noticeable difference (JND) of DRR as a lower bound to audiovisual mismatches: If a difference is not perceivable at all, it cannot lead to incoherence either. In one of our preceding experiments, we estimated the JND. The thresholds obtained ranged from 4–10 dB when level cues were not available, and approximately 2 dB when level cues were present. Thus, a noteworthy difference between JND experiments and the present audiovisual coherence experiment is the importance of roving, which is critical for JND tests, but does not seem to have a large influence on the presented coherence design.

5 Conclusion

In this paper, we presented an experiment that tests audiovisual coherence by varying DRR in an AR holographic calling scenario. As expected, participants were most likely to accept an audiovisual presentation as coherent if the DRR was matched to that of a reference rendering based on a measurement of the real room. When mismatches were created, the proportions of presentations perceived as coherent decreased; 50% thresholds were found to be approximately 3–4 dB on average. There was no evidence for a significant difference between the presentation of a talking audiovisual avatar and a loudspeaker emitting noise in the present study. Level roving also did not appear to have a substantial impact on audiovisual coherence in our experiment. Yet, to provide further evidence for the absence of these effects, a larger study would be required. Moreover, audiovisual coherence only presents one possible aim for AR audio rendering. In the future, results from DRR coherence experiments should be compared to results obtained in other paradigms: it could be tested how the threshold for creating an auditory illusion relates to the coherence threshold. Moreover, experiments assessing broader quality aspects could be conducted to determine whether cases exist in which the presentation is perceived as coherent, but the overall quality suffers from the mismatch regardless.

Acknowledgments

This work was supported in part by the European Union under Grant 101070672 (SPIRIT – Scalable Platform for Innovations on Real-Time Immersive Telepresence Project, https://www.spirit-project.eu/). The authors also acknowledge the use of HUCE infrastructure of the Aalto School of Electrical Engineering.

References

[1] Neidhardt, A., Schneiderwind, C., and Klein, F., "Perceptual Matching of Room Acoustics for Auditory Augmented Reality in Small Rooms - Literature Review and Theoretical Framework," *Trends Hear.*, 26, 2022, doi:10.1177/23312165221092919.

[2] Deppisch, T., Meyer-Kahlen, N., and Garí, S. V. A., "Blind Identification of Binaural Room Impulse Responses from Smart Glasses," *IEEE Trans. Audio Speech Lang. Process.*, 32, 2024.

[3] Xu, T., Li, J., Zu, P., Sahay, P., Kim, M., Obeng-Marnu, J., Miller, F., Qian, X., Passarella, K., Rachumalla, M., Nongpiur, R., and Shin, D., "Enhancing XR Auditory Realism via Multimodal Scene-Aware Acoustic Rendering," in *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology*, ACM, Busan Republic of Korea, 2025, doi:10.1145/3746059.3747730.

[4] Ratnarajah, A., Ghosh, S., Kumar, S., Chiniya, P., and Manocha, D., "AV-RIR: Audio-Visual Room Impulse Response Estimation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, doi:0.1109/CVPR52733.2024.02565.

[5] Brinkmann, F., Lindau, A., and Weinzierl, S., "On the authenticity of individual dynamic binaural synthesis," *J. Acoust. Soc. Am.*, 142(4), pp. 1784–1795, 2017, doi:10.1121/1.5005606.

[6] Meyer-Kahlen, N., Amengual Garí, S. V., Schlecht, S. J., and Lokki, T., "Testing Auditory Illusions in Augmented Reality: Plausibility, Transfer-Plausibility and Authenticity," *J. Audio Eng. Soc.*, 72(11), pp. 797–812, 2024, doi:10.17743/jaes.2022.0178.

[7] Lindau, A. and Weinzierl, S., "Assessing the Plausibility of Virtual Acoustic Environments," *Acta Acust.*, 98(5), pp. 804–810, 2012, doi:10.3813/AAA.918562.

[8] Schneiderwind, C., De Sena, E., and Neidhardt, A., "Perceptual effects of modified late reverberation and reverberation time in auditory augmented reality in two rooms," *Acta Acust.*, 10, 2026, doi:10.1051/aacus/2026012.

[9] Neidhardt, A. and Schneiderwind, C., "The influence of the DRR on audio-visual coherence of a real loudspeaker playing virtually over headphones," in *DAGA–Fortschritte der Akustik*, Vienna, Austria, 2021.

[10] Zahorik, P., Brungart, D. S., and Bronkhorst, A. W., "Auditory distance perception in humans: A summary of past and present research," *Acta Acust. united Ac.*, 91(3), pp. 409–420, 2005, doi: 10.3758/s13414-015-1015-1.

[11] Kopčo, N. and Shinn-Cunningham, B. G., "Effect of stimulus spectrum on distance perception for nearby sources," *J. Acoust. Soc. Am.*, 130(3), pp. 1530–1541, 2011, doi:10.1121/1.3613705.

[12] Kolarik, A. J., Moore, B. C. J., Zahorik, P., Cirstea, S., and Pardhan, S., "Auditory distance perception in humans: a review of cues, development, neuronal bases, and effects of sensory loss," *Atten., Percept., & Psychophys.*, 78(2), pp. 373–395, 2016, doi:10.3758/s13414-015-1015-1.

[13] Lavandier, M. and Prud'homme, L., "Reverberation cues underlying virtual auditory distance perception for a frontal source," *J. Acoust. Soc. Am.*, 158(3), pp. 1947–1961, 2025, doi:10.1121/10.0039251.

[14] Hládek, v., Seitz, A. R., and Kopčo, N., "Auditory-visual interactions in egocentric distance perception: Ventriloquism effect and aftereffect," *J. Acoust. Soc. Am.*, 150(5), pp. 3593–3607, 2021, doi:10.1121/10.0007066.

[15] Zahorik, P., "Asymmetric visual capture of virtual sound sources in the distance dimension," *Front. Neurosci.*, 16, 2022, doi:10.3389/fnins.2022.958577.

[16] Hofmann, A., Meyer-Kahlen, N., and Schlecht, S. J., "Audiovisual Congruence and Localization Performance in Virtual Reality: 3D Loudspeaker Model vs. Human Avatar," *J. Audio Eng. Soc.*, 72(10), 2024.

[17] Stenzel, H. and Jackson, P. J. B., "Perceptual thresholds of audio-visual spatial coherence for a variety of audio-visual objects," in *AES Conference on Audio for Virtual and Augmented Reality*, Redmond, WA, USA, 2018.

[18] Meyer-Kahlen, N., de las Heras, S., and Lokki, T., "Expected Levels of Reproduced Speech," in *AES Conference on Augmented and Virtual Reality*, Redmond, WA, USA, 2024.

[19] Acoustics Research Institute, "ARI HRTF Database. HRTFs with anthropometric data," `https://www.oeaw.ac.at/isf/outreach/software/hrtf-database`, 2026, Accessed: 2026-05-24.

[20] Bau, D., Moschner, O., and Pörschmann, C., "Adapting the PIRATE Design for Cost-Effective In-Ear Microphones," in *DAGA–Fortschritte der Akustik*, Hannover, Germany, 2024.

[21] Denk, F., Brinkmann, F., Stirnemann, A., and Kollmeier, B., "The PIRATE: an anthropometric earPlug with exchangeable microphones for Individual Reliable Acquisition of Transfer functions at the Ear canal entrance," in *DAGA–Fortschritte der Akustik*, Rostock, Germany, 2019.

[22] Mülleder, A., Romanov, M., Meyer-Kahlen, N., and Zotter, F., "Do-it-yourself headphones and development platform for augmented-reality audio," in *AES Conference on Immersive and Spatial Audio*, 2023.

[23] Pescoller, H., Mosleh, S., and Zotter, F., "Prototyping Interactive 3D Audio Applications in Pure Data with Multi- channel Connections and mtx_convolver~," in *33rd Tonmeistertagung – VDT International Convention*, 2025.

[24] Depthkit, "Unity Expansion Package," 2025, https://www.depthkit.tv/unity-expansion-package, Accessed: 30 Mar 2026.

[25] Schütt, H. H., Harmeling, S., Macke, J. H., and Wichmann, F. A., "Painfree and accurate Bayesian estimation of psychometric functions for (potentially) overdispersed data," *Vis. Res.*, 122, pp. 105–123, 2016, doi:10.1016/j.visres.2016.02.002.

A Appendix

Figure 6 shows the individual results for all 12 participants.

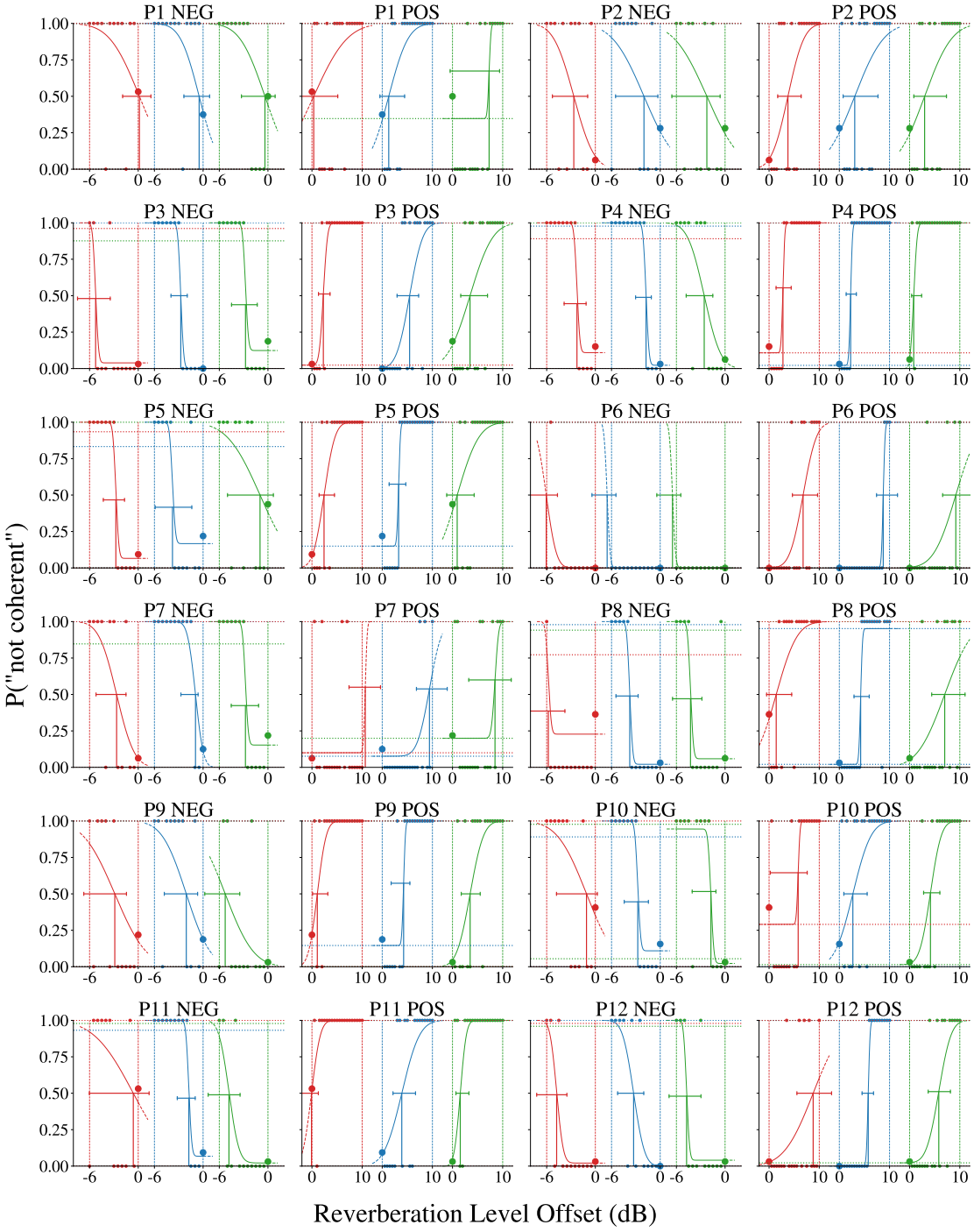


Fig. 6: Psychometric functions fitted to individual participants. Red indicates the speech condition, blue indicates the ±3 dB roving noise condition, and green indicates the ±8 dB roving noise condition. The corresponding coloured dots indicate the participants' binary responses in each condition.