DOI: 10.61782/fa.2025.0962

VISION-TO-AUDIO VS AUDIOVISION-TO-AUDIO MATCHING SHOWS THE DIFFERENCE BETWEEN VIRTUAL ACOUSTICS FOR VR AND AR

Nils Meyer-Kahlen* **Lisha Qiu** **Tapio Lokki** **Johannes M. Arend**
Acoustics Lab, Department of Information and Communications Engineering,
Aalto University, Finland

ABSTRACT

Virtual reality (VR) experiences immerse users in a virtual world distinct from their own environment. In contrast, augmented reality (AR) applications, such as AR telepresence, add virtual objects to the users' real environment through semi-transparent displays or visually occluding devices with camera-based passthrough. At first sight, the necessary virtual acoustic technologies for both applications appear similar—position and orientation-dependent binaural room acoustic rendering. However, the perceptual demands on the rendering systems differ considerably. In VR, users can only compare an acoustic rendering to the visual scene they are transported to. In AR, however, real-world sound sources may coexist with virtual ones, allowing for auditory comparison between real and virtual sound. This study compares subjects' ability to match binaural speech rendering using different rooms to a given reference room presented visually, through another sound source rendered in it, or through both. Therein, the hypothesis is that subjects will be better at matching when an auditory reference is provided. This experiment illustrates a critical distinction between acoustic rendering for VR and AR.

Keywords: Audio for VR and AR, Binaural Audio, Virtual Acoustics, Spatial Audio, Multimodal Perception

1. BACKGROUND

Whereas virtual reality (VR) experiences immerse users in a virtual world distinct from their own environment, augmented reality (AR) applications add virtual objects to the users' real environment. This is achieved through special head-mounted displays (HMDs) that are either semi-transparent or have cameras and visually occluding screens that allow for visual passthrough.

Whether the acoustic rendering of virtual objects is suitable for VR or AR can be assessed in different ways, for example by quality ratings [1], or by checking whether the rendering can cause auditory illusions [2, 3]. Therein, an auditory illusion is said to take place if listeners believe a virtual sound source to be real. For either test, it can be assumed that the outcome partly depends on the perceptual match between the room acoustic rendering applied to the virtual object and the user's expectations towards how the object should sound in the given room [4]. In this contribution, we highlight that the available information for forming such expectations differs between VR and AR, potentially leading to different perceptual requirements for the spatial audio rendering.

In VR, an acoustic rendering might be perceived as matching to the virtual environment if it aligns with the user's acoustic expectations formed from exposure to the visual representation of the virtual room; the user consciously or subconsciously performs vision-to-audio (V-to-A) matching. Only a few studies have been published regarding V-to-A matching, still leaving many questions open [4]. In a recent study, users were asked to adjust the reverberation time (RT) and the room volume of an acoustic simulation to match a room model presented visually in VR [5]. Participants chose higher RTs and larger acoustic room sizes for visually larger rooms, but there was considerable inter-subject variability. In another study [6], participants were asked to scale a visual model of a concert hall according to their expectation of room size inferred from an acoustic simulation, and vice versa. Again, a dependence of RT and room size was found. Other exper-

**Corresponding author*: *nils.meyer-kahlen@aalto.fi.*

iments tested whether the position within a room can be matched between visual and acoustic rendering. Without dedicated room acoustic training, this proved to be a very difficult task for test participants [7], even when listening with their own ears in a so-called locoscope test [8].

A related type of experiment does not ask subjects to match visual and auditory renderings directly, but rather assesses the influence of a mismatch between the two on other outcome variables. In [9], for example, it was shown that a mismatched audiovisual presentation leads to lower scores in a presence questionnaire. Furthermore, it was shown that such a mismatch leads to lower externalization ratings for the presented audio [10]. This effect has been called the room divergence effect [11].

In AR, users will not only see the room they are in, but real-world sound sources may exist alongside the virtual ones, allowing for auditory comparison between real and virtual sources. Crucially, since real and virtual sources will usually be different and emit different signals, only a certain degree of acoustic comparability is given. Nevertheless, AR allows audiovisual-to-audio (AV-to-A) matching.

In this study, we compare subjects' ability to perform V-to-A and AV-to-A matching. To gauge the relative importance of audio and vision in AV-to-A matching, we include an audio-to-audio (A-to-A) matching task, where no room is shown visually. We hypothesize that both AV-to-A matching and A-to-A matching are easier than V-to-A matching. This would imply that creating a perceptually matching rendering is harder for AR than for VR. Note that we conduct the experiments using a VR headset and binaural rendering with headphones, so they focus only on the matching task and not on creating auditory illusions, which would make the requirements for AR rendering even more demanding. In the following section, we describe the rendering approach, the experimental design, and the three tested conditions in detail.

2. METHODS

2.1 Room Acoustic Measurements and Processing

The spaces selected for the study are 22 rooms in the "A-Grid" building of Aalto University, where the Acoustics Lab is located (Otaakari 5, Espoo, Finland). In each of these rooms, spatial room impulse responses (SRIRs) were measured at several source-receiver positions, using a GRAS VI-50 intensity probe (GRAS, Holte, Denmark) consisting of six omnidirectional capsules, and a Genelec 8331 loudspeaker (Genelec, Iisalmi, Finland), directed towards the microphone array. The source-receiver distance of the SRIR selected for the experiment was 1.5 m in each room. The rooms included seven different stairwells/hallways, six meeting rooms, five offices, two lecture halls, a kitchen, and a bathroom. For an overview of the rooms' acoustic properties, see Fig. 1.

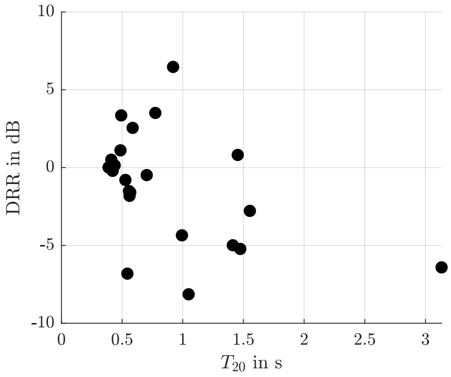
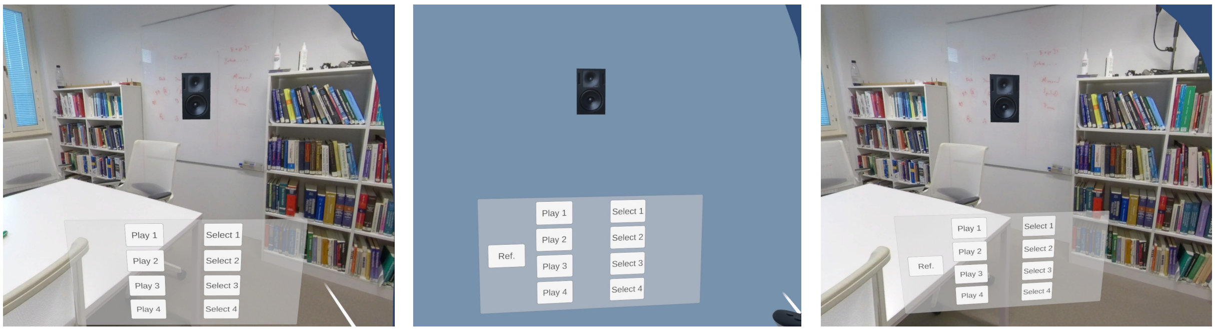


Figure 1. Mid-frequency (average of the 500 Hz and 1 kHz octave bands) reverberation time (T_{20}) and direct-to-reverberant energy ratio (DRR) of the 22 rooms used in the study.

For the experiment, third-order Ambisonics signals were created based on the measured microphone array SRIR using the spatial decomposition method (SDM) [12]. Therefore, the direction of arrival (DoA) is determined for each time instance of the response using time difference of arrival estimation. Then, according to these estimates, each sample of the omnidirectional responses was encoded to third-order Ambisonics, as proposed in [13]. Moreover, the DoAs of each response were rotated before encoding such that the direct sound always arrived exactly from the front.

2.2 Anechoic Speech

As speech material, anechoic sentences were selected from a dataset of anechoic speech of 21 speakers [14]. The material was recorded in a large anechoic chamber at a distance of 1.5 m from the speaker. The natural level differences between speakers were not normalized for the present experiment. 17 speech samples in nine different languages were selected from the "native" part of the dataset, for which participants were asked to record adhoc translations of Harvard sentences to their native lan-



(a) (b) (c)

Figure 2. Screenshots from the HMD during the experiment. Loudspeaker model seen in the middle, user interface in the foreground. (a) Vision-to-audio condition; the room is shown, but no reference audio playback is included. (b) Audio-to-audio condition; no 360° picture of the room is shown, but now the user interface contains an extra button labeled "Ref." on the left that plays the reference audio. (c) Audiovision-to-audio condition; visual room rendering and reference audio button are included.

guage. The reason for using different languages was to make listeners focus on the acoustics rather than on the content of the speech.

2.3 Visual Rendering

In addition to the SRIR measurements, stereoscopic 360° pictures were taken with an Insta 360 Pro 2 (Arashi Vision, Shenzhen, China) camera at the approximate receiver positions. The stereoscopic 360° pictures were processed and loaded into Unity as materials, so that they could be shown in an HMD. As stereoscopic 360° pictures provide depth information, it is possible to add additional objects to the picture as long as translational movements are disabled and the objects do not intersect with objects in the room. Using this method, a loudspeaker model was added in front of the subject at a distance of 1.5 m. Since the DoA data was rotated such that the direct sound was in the front, the loudspeaker model and the direct sound are precisely aligned, despite possible deviations in the microphone array and camera rotation.

2.4 Experimental Design

With the acoustic renderings and the pictures at hand, the listening experiment application was developed. It consisted of a control program written in Python, a player for head-tracked Ambisonics playback in Max MSP (Cycling'74) with the SPARTA ambiBIN [15] plugin using the MagLS decoder design [16], and a VR application developed in Unity and presented using an Oculus Quest 3 (Meta, Menlo Park, USA). The VR application also featured a simple user interface. The required audio and visual stimuli, head-tracking data, and responses were sent between applications using open sound control (OSC) messages. The audio was played back using Sennheiser HD 599 (Sennheiser, Wedemark, Germany) headphones. The following three conditions were implemented using this setup.

Vision-to-Audio (V-to-A, à la VR): A room including the loudspeaker model was shown in the HMD, see Fig. 2(a). Four auditory renderings using four different SRIRs could be played back by the participant by clicking on one of the four play buttons in the left column of the user interface. One of the responses belonged to the room shown visually, whereas the other three were randomly selected from the dataset. A different anechoic speech file, also randomly selected from the speech dataset, was used for each of the four renderings. The participant's task was to select the rendering that matches the room shown visually by clicking on the corresponding select button (right column of the interface). Hence, the experiment constitutes a 4-alternative forced-choice (4-AFC) paradigm.

Audio-to-Audio (A-to-A): In this condition, no room was shown, but a gray background was presented, see Fig. 2(b). Instead of a visual reference in the V-to-A condition described above, an audio reference rendering is now provided. Crucially, the reference speech was rendered in one of the four rooms presented as options, but a different speech file was used. As the listener has to transfer room acoustic properties from the reference to the four options in order to compare them, such a reference has been called "transfer" reference before [17]. It accounts for the fact that real and virtual objects are unlikely to emit the same signal.

Such a scenario would resemble an AR application, where participants cannot see the room. This is a less likely application case, and the condition is included for studying the relative importance of visual and auditory reference in the audiovisual-to-audio condition.

Audiovision-to-Audio (AV-to-A, à la AR): In this condition, illustrated in Fig. 2(c), the two previous conditions are combined; both the room was shown visually, and an acoustic "transfer reference" rendering was provided. This condition mimics an AR application where, in addition to seeing the room, a sound source might be heard that serves as an indirect reference.

The three conditions were presented in blocks, with the order counterbalanced across participants. In each block, all 22 rooms served as reference once, so that each of the three blocks consisted of 22 trials. Before the start of the study, participants gave their informed consent, and after it ended, they completed a short questionnaire on demographics and acoustical experience.

2.5 Participants

Twelve participants took part in the study. Most of them were students in the bachelor's program "Engineering Psychology" and some in the master's program "Acoustics and Audio Technology". Half of the participants indicated having some academic training in acoustics or audio. The mean age was 24.6 years (SD = 4.7 years), and none of the subjects reported any hearing loss.

3. RESULTS

The results of the study are presented in terms of percentage correct scores in Fig. 3. As the design represents a 4-AFC task, the guessing rate is $p_{\text{guess}} = 25\%$. The median responses are all clearly above the guessing rate. In the V-to-A condition, however, there were three participants who were close to guessing.

Comparing the three conditions, differences become apparent. The median percentage correct for the V-to-A condition is at 53.9%, whereas the correct response rate is almost 20% higher for the AV-to-A condition (71.8%). A signed rank test with Bonferoni-Holm correction for the three comparisons yielded a statistically significant difference between V-to-A and AV-to-A ($p = 0.004$).

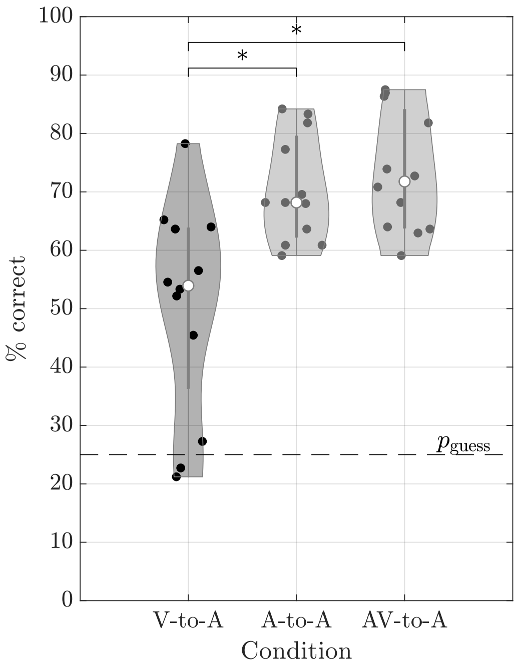


Figure 3. Percentages of correct responses obtained in the three conditions. The white dots show the median, and the stars signify statistically significant differences.

Interestingly, the percentage of correct answers in the A-to-A condition, where no image of the room was shown, is 68.2%, which is thereby much closer to the AV-to-A condition than to the V-to-A condition. The difference between visual and auditory matching (V-to-A vs. A-to-A) was also significant ($p = 0.007$). This suggests that pure acoustic matching is easier than matching the visual appearance of a room to the perceived room acoustics.

Fig. 3 also shows a much higher variance in percentage correct scores in the V-to-A condition than in other conditions. While some participants performed much worse in the V-to-A condition, others had similar levels of performance between the conditions. We did not find any correlation between performance and experience in acoustics research for any of the tasks.

4. DISCUSSION

The results clearly show that AV-to-A matching, which mimics an AR application, is much easier than V-to-A matching encountered in VR. Moreover, the difference between AV-to-A matching and A-to-A matching is very small compared to the difference between V-to-A matching and AV-to-A matching. This indicates that even an indirect acoustic comparison with another source signal provides much more information for forming expectations about a room than a visual presentation.

One of the main limitations of this study is that the transfer-reference and the four test stimuli were co-located, i.e., they all originated from the one presented loudspeaker. This does not represent a realistic scenario in an AR application, and the difference between co-located and non-co-located transfer-reference should be studied in a future experiment.

Also, instead of using two speech samples, albeit they were different from each other, much more dissimilar signals could be used to examine the matching. For example, the reference could be a short signal generated when entering a room, such as a few footsteps. Even one's own voice could be tested, even though this would be more challenging to implement.

Lastly, the study could be repeated in real rooms, in which participants are placed with or without the opportunity of listening to sounds in them, similar to how it was done in the locoscope test [8].

5. CONCLUSION

We have shown that correct perceptual audiovisual-to-audio room matching is significantly easier to complete than vision-to-audio matching, highlighting the strong difference between the requirements on acoustic rendering for VR and AR. Moreover, participants performed almost as well in an audio-to-audio matching task as in an audiovisual-to-audio matching task, indicating that having access to an acoustic reference in the room facilitates acoustic matching more than having a visual representation of the space.

6. REFERENCES

[1] F. Stärz, S. Van De Par, L. O. Kroczek, S. Roßkopf, A. Mühlberger, and M. Blau, "Comparison of binaural auralisations to a real loudspeaker in an audiovisual virtual classroom scenario: Effect of room acoustic simulation, HRTF dataset, and head-mounted display on room acoustic perception," *Acta Acustica*, Mar. 2025. Accepted.

[2] K. Brandenburg, S. Werner, F. Klein, and C. Sladeczek, "Auditory illusion through headphones: History, challenges and new solutions," in *22nd International Congress on Acoustics (ICA)*, (Buenos Aires, Argentina), Sept. 2016.

[3] N. Meyer-Kahlen, S. V. Amengual Garí, S. J. Schlecht, and T. Lokki, "Testing Auditory Illusions in Augmented Reality: Plausibility, Transfer-Plausibility and Authenticity," *J. Audio Eng. Soc.*, vol. 72, no. 11, pp. 797–812, 2024.

[4] A. Neidhardt, C. Schneiderwind, and F. Klein, "Perceptual Matching of Room Acoustics for Auditory Augmented Reality in Small Rooms - Literature Review and Theoretical Framework," *Trends in Hearing*, vol. 26, May 2022.

[5] B. Burnett, A. Neidhardt, Z. Cvetković, H. Hacıhabiboğlu, and E. De Sena, "User Expectation of Room Acoustic Parameters in Virtual Reality Environments," in *Immersive and 3D Audio: from Architecture to Automotive (I3DA)*, (Bologna, Italy), Sept. 2023.

[6] M. Frank and D. Perinovic, "Matching auditory and visual room size, distance, and source orientation in virtual reality," in *AudioMostly 2022*, (St. Pölten, Austria), pp. 80–83, Sept. 2022.

[7] F. Klein, A. Neidhardt, M. Seipel, and T. Sporer, "Training on the acoustical identification of the listening position in a virtual environment," in *143rd AES Convention*, (New York, NY, USA), Oct. 2017.

[8] N. Meyer-Kahlen, S. J. Schlecht, and T. Lokki, "Clearly audible room acoustical differences may not reveal where you are in a room," *J. Acoust. Soc. Am.*, vol. 152, pp. 877–887, Aug. 2022.

[9] P. Larsson, D. Västfjäll, P. Olsson, and M. Kleiner, "When What You Hear is What You See: Presence and Auditory-Visual Integration in Virtual Environments," in *Proceedings of the 10th Annual International Workshop on Presence*, (Barcelona, Spain), pp. 11–18, Oct. 2007.

[10] J. C. Gil-Carvajal, J. Cubick, S. Santurette, and T. Dau, "Spatial Hearing with Incongruent Visual or Auditory Room Cues," *Sci Rep*, vol. 6, Dec. 2016.

[11] S. Werner, F. Klein, T. Mayenfels, and K. Brandenburg, “A summary on acoustic room divergence and its effect on externalization of auditory events,” in *2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX)*, (Lisbon, Portugal), June 2016.

[12] S. Tervo, J. Pätynen, A. Kuusinen, and T. Lokki, “Spatial Decomposition Method for Room Impulse Responses,” *J. Audio Eng. Soc.*, vol. 61, pp. 17–28, Mar. 2013.

[13] M. Frank and F. Zotter, “Spatial impression and directional resolution in the reproduction of reverberation,” in *DAGA - Fortschritte der Akustik*, (Aachen, Germany), Mar. 2016.

[14] A. Hofmann, N. Meyer-Kahlen, and S. J. Schlecht, “Audiovisual Congruence and Localization Performance in Virtual Reality: 3D Loudspeaker Model vs. Human Avatar,” *J. Audio Eng. Soc.*, vol. 72, no. 10, pp. 679–690, 2024.

[15] L. McCormack and A. Politis, “SPARTA & COMPASS: Real-time implementations of linear and parametric spatial audio reproduction and processing methods,” in *AES International Conference on Immersive and Interactive Audio*, (York, UK), Mar. 2019.

[16] C. Schörkhuber, M. Zaunschirm, and R. Höldrich, “Binaural Rendering of Ambisonic Signals via Magnitude Least Squares,” in *DAGA - Fortschritte der Akustik*, (Munich, Germany), Mar. 2018.

[17] T. McKenzie, N. Meyer-Kahlen, and S. J. Schlecht, “The role of source signal similarity in distinguishing between different positions in a room,” in *AES Conference on Immersive and Spatial Audio*, (Huddersfield, UK), Aug. 2023.