ABILITY TO IDENTIFY CONCERT HALL ACOUSTICS BASED ON STEREOSCOPIC 360-DEGREE VISUALS

T Lokki Aalto Acoustics Lab, Dept of Information and Communications Engineering, Finland
L Qiu Aalto Acoustics Lab, Dept of Information and Communications Engineering, Finland
N Meyer-Kahlen Aalto Acoustics Lab, Dept of Information and Communications Engineering, Finland

1 INTRODUCTION AND MOTIVATION

Current Virtual Reality (VR) and multichannel sound reproduction technologies enable us to design audio-visual listening tests with stimuli captured in real concert halls. The equipment allows for a visual and aural experience that mimics a specific space in the laboratory, facilitating comparisons between existing concert halls. Earlier, our target was to identify all possible differences in room acoustics through rigorous listening tests using sensory evaluation methods[7,8,9]. These studies did not include visual cues. Thanks to researchers at the Institute for Hearing Technology at the RWTH Aachen, nowadays we also have realistic visual captures of some halls[3]; therefore, in this paper, we can report on a study in which participants saw stereoscopic 360° images on a Head-mounted Display (HMD) while they were listening to auralizations of concert halls in a multichannel anechoic room. The purpose was not to rigorously compare concert hall acoustics; rather, it was to test whether relatively naive participants could map the visual image to specific room acoustics. Specifically, the purpose was to find out whether a participant can match music rendered on one seat of a concert hall with a picture taken from that seat, and if music rendered in one concert hall can be associated with one of four concert halls presented visually.

1.1 Audio-visual studies on small room acoustics

Recently, we studied audio-visual perception in small rooms with a study[15], in which the participants had three different tasks:

1. *audio-to-video* matching; The participants saw a room in their HMD and could listen to four auralizations with headphones. The auralizations were realized with measured spatial room impulse responses (SRIR) convolved with four different speech signals. The task was to select an auralization that matched what a participant saw in their HMD. The success rate was about 55%, thus not very high but clearly above the guessing rate of 25%.
2. *audio-to-audio* matching; In this task, there was an auralization of a room with speech, and the participants had to select a rendering that was created in the same room out of four different alternatives. The reference rendering used a different speech signal than the four auralizations. In other words, the listener had to transfer room acoustic properties from the reference to the four options in order to compare them. In studies regarding small room acoustics, this was also called "transferring task"[12,14]. Audio-to-audio matching using different signals was also used in an earlier study regarding concert hall acoustics[6]. In our recent test, the participants achieved 70% correct in these small rooms, which indicates that people could recognize small room acoustics to some extent, when a different speech rendering was provided as the reference.
3. *audiovisual-to-audio* matching; This task was the same as the audio-to-audio case but with visuals. The results showed that participants did only slightly better than in the audio-to-audio case; thus, the visuals did not clearly increase the recognition rate. This result could also be interpreted so that visuals did not help at all, and participants just related to aural cues.

Another type of audiovisual matching study conducted in small rooms involves identifying the position within a room. In one study, 360° pictures from different positions of a room were shown and listeners had to associate a rendering to these positions[4]. Only some listeners could perform the task, and only after extensive training. We conducted a similar positional matching experiment, but employed a trick that allowed them to hear sound in a real room when they were asked to associate it with visual VR renderings from different positions in the room; again, the task was very difficult for subjects[16].

1.2 Audio-visual concert hall studies

In general, there are not many audio-visual studies on concert hall acoustics. Maybe the most advanced and realistic stimuli have been realized in TU Berlin by Maempel and Horn[10]. It consisted of a large visual display with a hologram video of a quartet playing, thus very realistic visuals. The binaural sound was based on binaural room impulse responses, captured with a movable dummy head enabling head-tracked binaural playback. Thus, the system was the state-of-the-art both in visual and sound quality. However, they have used it only in a few published studies[10,11], mainly on distance and room size perception.

Chen and Cabrera have used computer graphics to visualize the concert hall in different colors and combine those with binaural sound. The audio was produced by combining anechoic recordings with measured second-order Ambisonics room impulse responses[1]. They found small differences in sound perception depending on the colors of the hall. In addition, they studied the audio-visual seat preference[2] with simulated auralizations.

Recently, a study on the audio-to-vision matching of concert halls has been presented[18], in which pairs of acoustic renderings were compared to one visual rendering, and participants were asked to indicate which one they believed matched better. There was strong agreement between participants for some pairs, indicating consistent expectations about the acoustics of the space presented visually. However, the acoustic differences between the options were also rather drastic, ranging from anechoic to 4 s of reverberation time at mid frequencies. Moreover, there was no objectively correct rendering, as the visuals showed a virtually designed hall that does not exist in the real world.

1.3 Present Study

Building on the above findings, this study explores auditory spatial awareness in concert halls through an audio-to-vision matching paradigm. We test whether the participants can map visualized concert halls to what they hear. To this end, participants are listening to multichannel reproductions of concert halls while simultaneously viewing stereoscopic 360° images through an HMD. In the first part of the test, participants can switch between images taken at three different positions in the same concert hall, and they are asked to identify the seat corresponding to the reproduced audio. We expect this positional matching to work somewhat reliably, as it was shown that at least relative auditory distance judgments in pair-wise comparison are possible in concert halls, even though they depend on the hall[5]. In the second part, participants are asked to match the visuals of four different halls to an auralization. These visuals are all captured at the same distance from the orchestra, and an auralization corresponds to the same distance.

One purpose of this paper is to investigate whether it is possible at all to quickly compare concert halls if the participants cannot do rigorous AB comparison with the same signal over and over again. Based on our experience, when listening to the same signal attentively for more than 15 minutes, people start to hear even tiny differences, even though at first glance they report no differences between signals. Indeed, we have shown that recognition of acoustical features in concert halls is much harder when the listened signal is not exactly the same[6*]. Therein, many participants did not perform better than chance level, but they tended to confuse acoustically more similar halls with each other; two shoebox-

*A short binaural version is still available online: `anttikuusinen.shinyapps.io/RoomAcousticsMatchingExperiment/`

shaped halls were confused more frequently than a shoebox-shaped hall with a fan- or vineyard-shaped hall. While this was an audio-to-audio matching study, we expect to see similar patterns in the presented audio-to-video matching experiment.

2 METHODS

2.1 Visual and acoustic stimuli

Using the available multichannel auralizations and stereoscopic 360° images, a listening experiment application was created. This application comprised a control program developed in Python, a Max MSP-based player for auralization, and a VR application built in Unity, presented on an Oculus Quest 3 headset. The VR component also included a basic user interface. Open Sound Control (OSC) messages were used to transfer necessary metadata between the different software and to collect the answers of the participants.

360° stereoscopic photographs were taken in four concert halls[3†]; the Cologne Philharmonie (CP), the Philharmonic hall in Munich Gasteig (MG), Lahti Sibelius Hall (LS), and Beethovensaal in Stuttgart Liederhalle (SB). The camera was a Kandao Obsidian S, placed at a height from the floor ranging between 105 and 120 cm. Its settings were adjusted to manage brightness without altering color; each room had a reference photograph to which all others were matched, ensuring consistent luminance and color. Individual lens images were stitched into a 360° view and stored as .jpg files that measured 8760×2880 pixels. For VR display, the bitrate was raised from 24 to 32, the resolution was upscaled to 15360×8640 pixels, and it was saved in .hdr format. In Unity, a 3DoF experience was designed that shows the stereoscopic 360° images, creating the impression of being in a concert hall.

The auralizations were created with the loudspeaker orchestra measurements, see e.g.[8], rendered with anechoic symphonic music[17]. The music excerpts were 8 seconds long and played repeatedly while the participant was listening to them and comparing visuals. To avoid learning effects, we used eight different music clips (from Beethoven, Bruckner, and Mahler), randomly selected for each task. We considered randomizing the musical excerpts to be important to prevent the participants from comparing the task with the previous one, as we have found that participants often take the previous sample as a reference, e.g., in terms of loudness. The played music excerpts maintained their original level differences, the reproduction levels being from 63 to 84 dB (L_{Aeq}). The audio playback used 45 Genelec 8331 loudspeakers in the Aalto acoustics lab anechoic multichannel room.

2.2 Listening test

The listening test had two parts, one related to the recognition of a seat within a concert hall and another one for matching an auralization of a hall to one of four different visually presented halls. There was no training before the first test, but it was made clear that the participants understood the tasks and that they were able to use the controller and operate with the user interface, which is seen in screen captures in Fig. 1.

The first test was a **Position test** in which a participant heard an auralization of a concert hall and saw a 360° stereoscopic image of a concert hall. With a controller, the participant could switch the viewpoint between the three different seats, and the task was to select the seat that corresponded to the sound. The seats were positioned slightly off the central axis of the hall, at distances of 11 and 23 meters, and one seat further away, typically on the side. Thus, potential acoustic cues for selecting the correct seat were direct-to-reverberant ratio (DRR), loudness, and clarity. In total, there were 22 tasks, as 11 seats (one hall had auralizations only from two seats) were tested twice. To avoid learning effects, the order of auralizations of different seats was randomized. In addition, for every task, the different music sample was randomly selected among eight different samples, as described above.

[†]Available at: https://doi.org/110.5281/zenodo.15409976

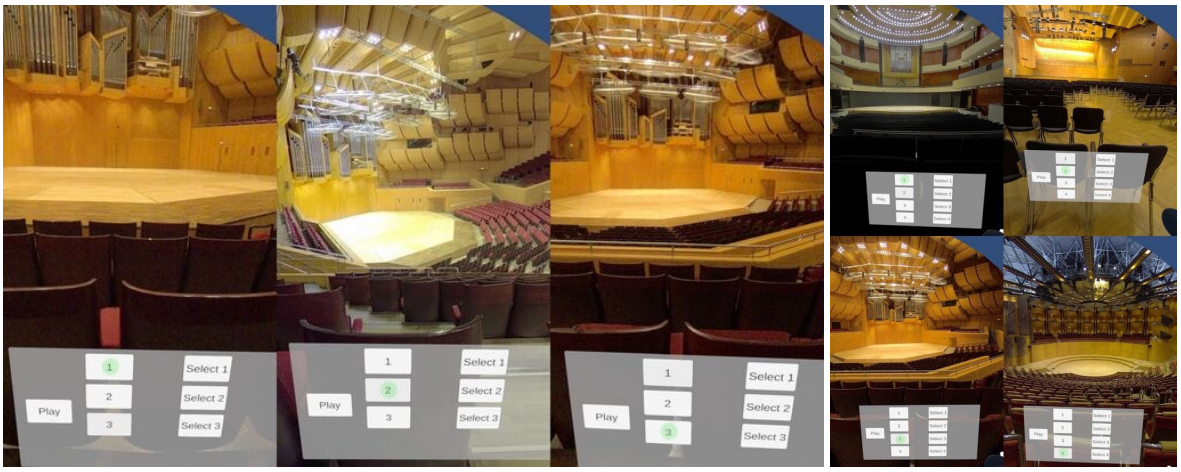


Figure 1 Screenshots from the HMD during the experiment. On the left: This shows the first test. Participants listened to an auralization, and they were shown VR visuals of three positions within the same music hall. Their task was to select the visual that corresponds to the auralization. The green dot indicates the currently viewed visual. On the right: This shows the second test. Participants listened to an auralization, and they were shown VR visuals at the same positions in four different concert halls. They needed to switch between the visuals of the four halls and were asked to select the one that corresponded to the auralization.

The second test was a **Room recognition test**. Again, a participant heard one auralization (in random order and with a random music sample) and she could switch between visuals, captured at the same seat, of four different concert halls. The task was to select the one which she thought corresponded to the auralization. Eight seats (auralizations at 11 or 23 meters in 4 concert halls) were tested twice, resulting in a total of 16 tasks.

2.3 Participants

Fourteen participants, mainly students in acoustics (none of the authors participated), took part in the study. They reported having between zero and three years of experience in acoustics research (mean = 1.19 years, SD = 1.03 years). The mean age was 27.7 years (SD = 4.2 years) and none of them reported any hearing loss. The post-test questionnaire revealed that the participants considered the tasks in both tests to be hard. They did not name any common strategy to recognize the halls; however, some participants mentioned they used cues such as distance, loudness, width of the orchestra, direction, and reverb.

3 RESULTS

The main result of the tests is plotted in Fig. 2. As the tests were repeated, we plot here all 14 participants twice (both ratings) resulting in a total of 28 data points for both tests. As can be seen, the participants were not really good at mapping the heard acoustics to the images. In the Position test, the median is above the guessing level, and some participants did quite well with over 2/3 correct answers. In contrast, the mapping of heard acoustics to one of the four halls in the Room recognition test cannot be distinguished from guessing. A few participants answered more than 50% of tasks correctly, but almost as many were even below the guessing level.

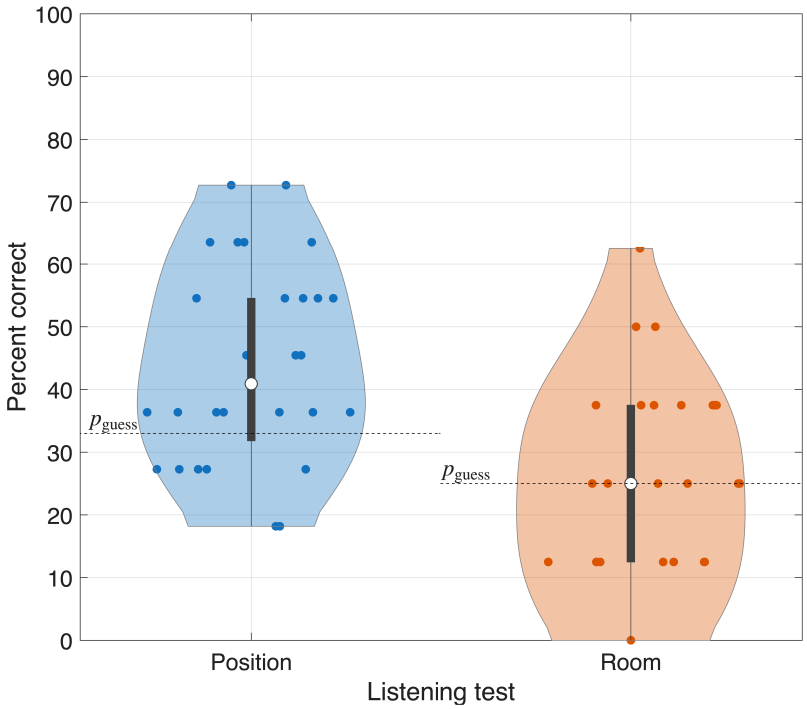
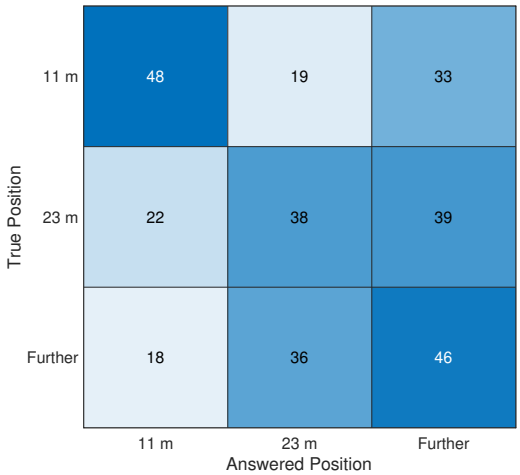
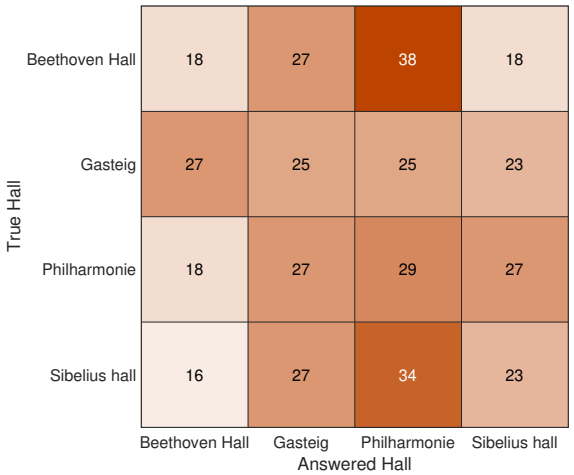


Figure 2 The distribution of percentage of correct answers for each participant. The white dots represents the median and the black bars show the 25 to 75% range of the data.



(a) Percentage of correct results in the Position test (guessing level: 33%).



(b) Percentage of correct results in the Room recognition test (guessing level: 25%).

Figure 3 Confusion matrices of both tests. Values show the percentage of correct answers.

The confusion matrices shown in Fig. 3 would reveal if some of the seats or halls were easier for participants. As expected, the closest seat at 11 meters was confused less with more distant seats. The farther seats were confused more with each other, which was also expected. Yet, all in all, the percentages are low, considering the fact that the heard music had to be mapped to one of only three seats (guessing level 33%). The low performance can partly be attributed to the fact that the music and its level changed for each task, so that no direct loudness cue was available, and the main cues for participants were probably DRR and clarity. The other confusion matrix in Fig. 3 confirms the result that the participants were mostly guessing. There is no real trend seen in the confusion matrix.

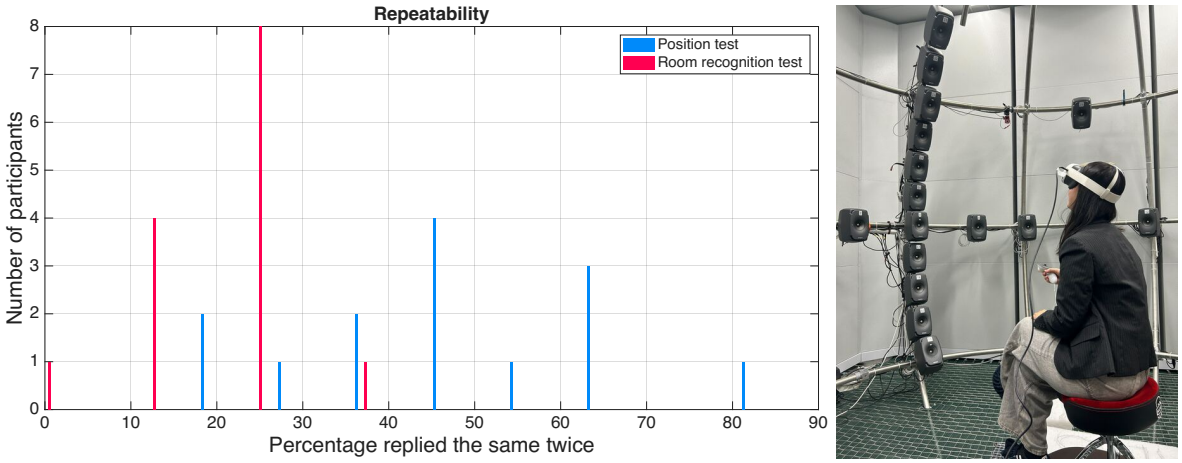


Figure 4 On the left: Percentage of how often the participants answered the same position/hall in the listening test, which had one replication for each case. On the right: a participant in the listening test.

The repetition of all stimuli allows us to check the consistency of the participants. In the analysis, it was checked if a participant answered exactly the same image twice, regardless of whether the answer was correct or wrong. Figure 4 shows the histogram of percentages of repeatability. In the Position test, on average, the participants could repeat the answers with 48% accuracy. The best participant could replicate the choices at an 82% rate, and as can be seen, there was quite a lot of variance between participants. In the Room recognition test, results show that it was really hard as the average was only 21% which is around the guessing level, as the main result already indicated. To conclude, none of the participants did very well.

Finally, we checked for correlations between the years of experience in acoustics research and task performance. For the Position test, there was no significant correlation ($R^2 = 0.02, p = 0.61$). For the Room recognition test, however, there was ($R^2 = 0.46, p = 0.01$). Accordingly, the best-performing participant had the most experience, and the worst-performing participant had the least.

4 DISCUSSION

Overall, the participants performed at the guessing level when they needed to map an auralization to one of the four visualized halls. This result is not entirely unexpected; it highlights that being able to map the seen environment to room acoustics is purely an expert ability, and it cannot be done without sufficient acoustics training. This interpretation is reflected by the significant correlation between performance and years of experience.

The Position test was comparatively easier, and particularly the frontal seat was recognized more often. However, we expected much higher recognition rates. The usage of eight different music excerpts prevented learning the differences between seats, as the overall loudness varied from task to task, which is different from previous distance experiments in concert halls [5]. Being able to make use of the remaining loudness differences would require having rather precise expectations of the absolute loudness of an orchestra in a specific passage. It has been shown that approximate absolute loudness expectations exist for speech [13], but no studies exist for orchestral music yet.

After obtaining the described results, the question arose whether performance would be better if the matching task was reversed; so that instead of selected a visual to match the presented audio rendering, participants would select an audio rendering to match the presented visual. We expected this to lead to better performance, as the more immediate comparison of subtle acoustic differences

might be beneficial—particularly in the Position test. As a follow-up to the original experiment, we implemented a revised version and conducted it with five participants (mean age = 28.6 years; mean experience in acoustics research = 2.6 years), none of whom had participated in the initial test. However, the resulting percentages of correct responses were not higher. In both parts of the test, they were approximately 29%. For the Position test, this places the results at the lower end of the previously observed range. Again, different orchestra samples were used, which suggests even being able to play the audio for the three positions back to back does not improve performance when participants lack sufficient absolute level expectations to account for level differences between the samples.

5 CONCLUSION

The results of this study highlight the complexities involved in accurately mapping auditory stimuli to visual representations in concert halls. The experiments conducted demonstrate that participants struggled in associating the acoustics of a particular hall with its corresponding visual representation in the Room recognition test, basically performing at chance levels. Despite incorporating state-of-the-art technologies like multichannel auralizations and 360° stereoscopic visuals, the task proved to be challenging for participants, suggesting that specialized training or a more consistent auditory framework might be necessary to improve performance in similar future experiments.

In the Position test, participants performed slightly better at identifying seats within a single concert hall based on auditory cues. However, the anticipated higher recognition rates were not achieved, likely due to the variability in music excerpts, which reduced reliance on loudness as a distinguishing auditory cue. This suggests that absolute loudness expectations depending on the music, and the usable auditory cues, such as direct-to-reverberation ratio and clarity, are not so easy to interpret without extensive training.

Overall, this study provides valuable insights into audio-visual listening tests. It underscores the necessity for well-thought-out experimental designs that perhaps integrate more rigorous training protocols or implement more uniform sensory stimuli. This study also raised the question of whether audio-visual studies would bring any novelties to the research on concert hall acoustics.

Acknowledgments: The authors would like to thank Josep Llorca-Bofí and Jonas Heck for visual captures of the concert halls.

6 REFERENCES

1. Y. Chen and D. Cabrera. The effect of concert hall color on preference and auditory perception. *Applied Acoustics*, 171:107544, 2021.
2. Y. Chen, D. Cabrera, and D. Alais. Modelling audiovisual seat preference in virtual concert halls. *Applied Acoustics*, 212:109589, 2023.
3. J.Heck, J. Llorca-Bofí, H. Park, N. Meyer-Kahlen, T. Lokki, and M. Vorländer. Comparison of the auditory and auditory-visual perception in four concert halls. *Proceeding of meetings on acoustics*, 56, 2025. in press.
4. 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, October 2017.
5. A. Kuusinen and T. Lokki. Investigation of auditory distance perception and preferences in concert halls by using virtual acoustics. *J. Acoust. Soc. Am.*, 138(5):3148–3159, November 2015.
6. A. Kuusinen and T. Lokki. Recognizing individual concert halls is difficult when listening to the acoustics with different musical passages. *Journal of the Acoustical Society of America*, 148(3):1380–1390, September 2020.
7. T. Lokki, J. Pätynen, A. Kuusinen, and S. Tervo. Disentangling preference ratings of concert hall acoustics using subjective sensory profiles. *Journal of the Acoustical Society of America*, 132(5):3148–3161, November 2012.
8. T. Lokki, J. Pätynen, A. Kuusinen, and S. Tervo. Concert hall acoustics: Repertoire, listening position and individual taste of the listeners influence the qualitative attributes and preferences. *Journal of the Acoustical Society of America*, 140(1):551–562, July 2016.
9. T. Lokki, J. Pätynen, A. Kuusinen, H. Vertanen, and S. Tervo. Concert hall acoustics assessment with individually elicited attributes. *Journal of the Acoustical Society of America*, 130(2):835–849, August 2011.
10. H.-J. Maempel and M. Horn. The virtual concert hall - A research tool for the experimental investigation of audio-visual room perception. *International Journal on stereo & immersive media*, 1(1):78–98, 2017.

11. H.-J. Maempel and M. Horn. The influences of hearing and vision on egocentric distance and room size perception under rich-cue conditions. In Brian F.G. Katz and P. Majdak, editors, *Advances in Fundamental and Applied Research on Spatial Audio*, chapter 4. IntechOpen, Rijeka, 2022.
12. 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, 2023.
13. N. Meyer-Kahlen, S. de las Heras, and T. Lokki. Expected Levels of Reproduced Speech. In *AES Conference on Augmented and Virtual Reality*, Redmond, WA, USA, August 2024.
14. N. Meyer-Kahlen, S. V. Amengual Garí, I. Ananthabhotla, and P. Calamia. A Two-Dimensional Threshold Test for Reverberation Time and Direct-to-Reverberant Ratio. In *Immersive and 3D Audio: from Architecture to Automotive (I3DA)*, Bologna, Italy, September 2023.
15. N. Meyer-Kahlen, L. Qiu, T. Lokki, and J. Arend. Vision-to-audio vs audiovision-to-audio matching shows the difference between virtual acoustics for VR and AR. In *Forum Acusticum*, Malaga, Spain, June 2025.
16. N. Meyer-Kahlen, S. J. Schlecht, and T. Lokki. Clearly audible room acoustical differences may not reveal where you are in a room. *Journal of the Acoustical Society of America*, 152(8):877–887, August 2022.
17. J. Pätynen, V. Pulkki, and T. Lokki. Anechoic recording system for symphony orchestra. *Acta Acustica united with Acustica*, 94(6):856–865, November/December 2008.
18. P. Rolkowski, P. Odya, and B. Mróz. Reverberation divergence in VR applications. *International Journal of Electronics and Telecommunications*, 70(2):373–380, June 2024.