Meyer-Kahlen, Nils; Schlecht, Sebastian; Lokki, Tapio

Clearly audible room acoustical differences may not reveal where you are in a room

Published in:
Journal of the Acoustical Society of America

DOI:
10.1121/10.0013364

Published: 05/08/2022

Document Version
Publisher's PDF, also known as Version of record

Please cite the original version:
Meyer-Kahlen, N., Schlecht, S., & Lokki, T. (2022). Clearly audible room acoustical differences may not reveal where you are in a room. *Journal of the Acoustical Society of America*, *152*(2), 877-887. https://doi.org/10.1121/10.0013364

Clearly audible room acoustical differences may not reveal where you are in a room

Nils Meyer-Kahlen, Sebastian J. Schlecht and Tapio Lokki

Citation: The Journal of the Acoustical Society of America **152**, 877 (2022); doi: 10.1121/10.0013364
View online: https://doi.org/10.1121/10.0013364
View Table of Contents: https://asa.scitation.org/toc/jas/152/2
Published by the Acoustical Society of America

JASA ARTICLE

Clearly audible room acoustical differences may not reveal where you are in a room[a)]

Nils Meyer-Kahlen,[b)] Sebastian J. Schlecht,[c)] and Tapio Lokki
Acoustics Lab, Department of Signal Processing and Acoustics, Aalto University, P.O. Box 13100, FI-00076 Aalto, Finland

ABSTRACT:
A common aim in virtual reality room acoustics simulation is accurate listener position dependent rendering. However, it is unclear whether a mismatch between the acoustics and visual representation of a room influences the experience or is even noticeable. Here, we ask if listeners without any special experience in echolocation are able to identify their position in a room based on the acoustics alone. In a first test, direct comparison between acoustic recordings from the different positions in the room revealed clearly audible differences, which subjects described with various acoustic attributes. The design of the subsequent experiment allows participants to move around and explore the sound within different zones in this room while switching between visual renderings of the zones in a head-mounted display. The results show that identification was only possible in some special cases. In about 74% of all trials, listeners were not able to determine where they were in the room. The results imply that audible position dependent room acoustic rendering in virtual reality may not be noticeable under certain conditions, which highlights the importance of evaluation paradigm choice when assessing virtual acoustics.
 https://doi.org/10.1121/10.0013364
(Received 31 January 2022; revised 12 July 2022; accepted 14 July 2022; published online 5 August 2022)
[Editor: Brian F. G. Katz] Pages: 877–887

I. INTRODUCTION

"Can you hear where you are in a room?" is one possible question that may help to decide when the room acoustic rendering in virtual reality applications needs to be position dependent. In the first experiment presented, we assess the existence of audible position dependent acoustical differences in a room. In the second experiment, we test the ability to make sense of these differences using a newly developed virtual reality (VR) application, running on a head-mounted display (HMD). It allows for physically walking around in one zone within a room while creating the visual illusion of being in a different zone. Subjects are asked to associate the sound they hear within the zone they are physically located at with the correct visual renderings presented on the HMD. In other words, participants can examine the room acoustics with their own ears without any technology for sound rendering while switching between different visual renderings, one of which is correct. The room geometry and the five zones used in the test are shown in Fig. 1.

The design pays tribute to our view that the problem of locating oneself in a room is of multi-modal nature. In VR applications, direct comparison between virtual acoustic rendering and the sound of the specific source in the real world is not possible. To understand whether the virtual acoustic rendering is correct, users consciously or subconsciously need to infer acoustic features from the visual input or vice versa. The same is the case in the conducted experiment.

Before describing the audiovisual experiment, we first show the room acoustical differences in the tested room objectively and present the results of a paired comparison test. There, participants directly compared auralizations based on binaural room impulse responses (BRIRs), measured at the different positions in the room, without any visual cues.

After providing background on self-localization in a room and describing the overall study design in the following introduction, Sec. II describes the room in which the test was conducted in detail. Then, in Sec. III, the paired comparison test is presented. The audiovisual test is described in Sec. IV, and its results are presented in Sec. V. Then we discuss the results and limitations of the study in Sec. VI. Finally, in Sec. VII, we draw conclusions about the relevance of position dependent virtual acoustics in VR.

A. Background

Pioneering work on acoustically distinguishing different positions in a room was published by Shinn-Cunningham and Ram (2003), where participants had to recognize different positions in a room, reproduced based on static BRIRs. Already in this study, it became clear that determining one's own position in a room based on room acoustic information alone, i.e., when the location of sound sources is the same

[a)]The test design of the audiovisual experiment and partial results were presented in "Assessing room acoustic self-localization using a virtual blindfold," in *DAGA—Fortschritte der Akustik*, Vienna, Austria, August 2021.
[b)]Electronic mail: nils.meyer-kahlen@aalto.fi
[c)]Also at: Department of Art and Media, Aalto University, P.O. Box 13100, Aalto, Finland.

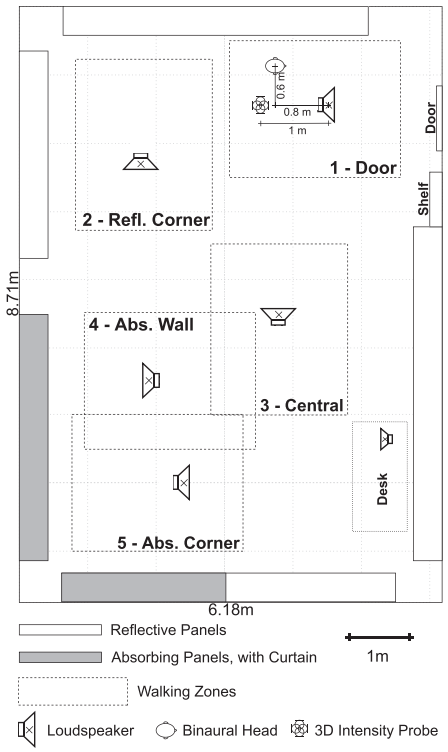

FIG. 1. Variable acoustics room Arni and the five loudspeaker zones. The left lower corner was set to an absorbing setting, and all other surfaces were reflective.

between all positions, is generally a difficult task. Participants were only able to differentiate two positions close to a wall from two positions without adjacent walls in some cases.

More recently, Neidhardt *et al.* (2016) conducted a number of related experiments. They used room simulations to conduct a test in which participants identified positions on a map, with static or dynamic binaural rendering, and using an omnidirectional or a directional loudspeaker as a virtual source. After a short training using a different signal, task performance was slightly above chance, and an improvement was observed in the course of the test. Identification was better for the omnidirectional source, and there was no effect of static vs dynamic synthesis. In Neidhardt (2016), emphasis was put on training. Participants were presented with pictures and a map of a meeting room along with dynamically rendered binaural sound. In a training session, they were allowed to add items to a single choice comparison one-by-one, according to their own liking and after examination of the individual items. Moreover, during the test, items were added consecutively so that participants first had to decide within a smaller subset. Feedback was also provided. Using this learning-focused approach, participants performed well above chance when selecting between a reduced set of stimuli and slightly above chance for the full set of five positions. In Klein *et al.* (2017), measurements from the same room were used along with 360° pictures presented over a HMD, and participants had to match one of four binaural reproductions to each picture. It was found that only some participants ("learners") were able to improve during extensive training. Therefore, it was concluded that without training, listeners would not notice position mismatches.

A different approach to the question of audibility and type of position dependent room acoustical differences was taken by Băcilă and Lee (2021). In their study, differences between acoustic attributes were assessed by comparing binaural renderings of different positions and orientations in a concert space in an elicitation study, based on paired comparisons. The three most salient attributes were connected to loudness, envelopment, and width. A listening test was also conducted *in situ*, by changing between different listening positions in the real hall. Similar attributes were elicited, with the main difference that fewer timbre-related attributes occurred. Here, the lack of direct comparison was seen as a shortcoming of such an *in situ* design, despite it being the closest to a VR application.

A highly related field of research is echolocation, reviewed, for example, by Kolarik *et al.* (2014). In several experiments, it was demonstrated that especially visually impaired people can develop excellent skills in using auditory cues to make sense of the surrounding space and their position within that space. The ability of blind and sighted subjects under various conditions has been studied, but the utilized acoustic differences underlying this performance are not completely clear. Kolarik *et al.* (2014) identify energy, spectral changes, binaural differences, and differences in the reverberation pattern as possible cues. They also highlight that to hear an echo "as such" is in any case only possible at long echo times beyond the echo threshold, which can be up to 50 ms, depending on the signal and other conditions (Litovsky *et al.*, 1999). At shorter echo times, the phenomenon called "time separation" pitch by Kolarik *et al.* (2014), i.e., comb-filtering, may play the largest role. In that sense, the term "echo location" can be misleading, as it should not refer to actually hearing the repeated sound, but more the consequences of it mixing with the direct sound. Whatever the strategies are, to employ them in an unknown room requires extensive training. Picinali *et al.* (2014), who conducted a study in which blind participants experienced a virtual acoustic environment, even mention that it is not uncommon for blind people to visit a new, unknown place like a new office building several times just to learn the room specific acoustic cues. Comb-filtering was also pointed out as one of the important cues when detecting the distance of a reflector based on measurements in an anechoic environment (Paasonen *et al.*, 2017). In this study, participants were asked to sort triplets of BRIR-based renderings by the proximity of a surface that was present during the measurement. For short distances of up to 27 cm, they were often able to do so correctly.

One refers to active echolocation when self-generated sounds are used, which is common when echolocation is performed by the visually impaired. The present study only examines passive echolocation, where external sounds can be used for orientation in the room.

https://doi.org/10.1121/10.0013364

B. Study design

All in all, prior work shows the existence of echolocation as an expert ability, on one hand, but suggests that for novice listeners without special training, associating the position in a room based on a map (Neidhardt, 2016; Neidhardt *et al.*, 2016; Shinn-Cunningham and Ram, 2003) or 360° pictures presented on a HMD (Klein *et al.*, 2017) with the acoustics is extremely hard. However, all experiments with novice listeners were done using headphone rendering, without individualized binaural synthesis, and they did not offer natural six degrees of freedom (6DoF) movements of the listener. If head-tracked playback was implemented [three degrees of freedom (3DoF)], participants were reported to rarely use it (Neidhardt *et al.*, 2016) or decreased their use after they had identified specific cues (Klein *et al.*, 2017).

To rule out the possibility that the bad performance of listeners who are not experts in echolocation is due to insufficient rendering quality, we implemented an audiovisual test design that allows participants to move freely while using their own ears to examine the sound in the room. It should be emphasized that our main focus lies in understanding if for a given room presented in VR the simulation of room reverberation needs to be listener position dependent. Therefore, as in other recent studies about quality aspects of virtual acoustic rendering for VR (Engel *et al.*, 2021; Lübeck *et al.*, 2022), we decided to use a group of sighted, normal hearing listeners without special experience in echolocation.

Before presenting the results of the audiovisual experiment, we describe a simple paired comparison test, in which participants were presented with renderings created from a BRIR measurement in each of the zones. The objective of this test was to determine whether the differences between the positions are audible in direct comparison. At the same time, participants were asked to identify the most salient acoustic attributes.

In the audiovisual test, wearing a HMD prevents participants from seeing the real room. Instead, an application developed for this experiment shows different rotated and shifted versions of the room. By doing so, it creates the visual illusion of being at a different position within it. Similar to an invertoscope, which creates the visual illusion of seeing the world upside-down, or a periscope, which allows for looking around a corner, we term this application *locoscope*. Ideally, the locoscope should be able to create strong visual place illusions (PIs) (Slater, 2009), evoking the sense of being at the shown places in the room. In the experiment, participants are then able to switch between different virtual positions and are asked to select the virtual position that they believe matches their position in the real room. A more detailed description is given in Sec. IV.

We hypothesize that participants will be able to find differences in the paired comparison test but will be close to guessing when identifying their position in the locoscope test, even though they are able to use their own ears and have unlimited response time. From this result, we could conclude that for the tested room, rendering of position dependent differences in room acoustics would be unnecessary for VR. This conclusion is naturally limited to rooms with a similar range of acoustical differences.

II. SELECTED ROOM

The variable acoustics room "Arni," described in more detail by Prawda *et al.* (2020), was selected for the study because it allows for strong acoustic variability within the room. Strong variability was desired such that acoustical differences between zones would be clearly audible. The walls are composed of 55 variable acoustic panels, which were all set to the reflective setting, except for those in the back left corner (see Fig. 1). These were set to the acoustically absorbing setting with a heavy curtain in front of them. The dimensions of the room, $8.7\,\mathrm{m} \times 6.18\,\mathrm{m} \times 3.5\,\mathrm{m}$, are in a similar range as the rooms used by Shinn-Cunningham and Ram (2003) and Neidhardt *et al.* (2016).

Five zones were defined in the room. In two zones, the loudspeaker played into the room (zones 2 and 3), with the difference that the sidewall and the backwall were more distant at the more central zone 3. Zone 4 was also relatively central, and it even overlapped with zone 3, but the loudspeaker was pointing toward the absorbing wall. In zone 1 on the other hand, the loudspeaker faced a strongly reflecting steel door, and participants were very close to a reflecting wall on the left. Finally, zone 5 was placed in the absorbing corner. The objective of the zone selection was again to create differences that would be clearly audible. Note that the location and the acoustic properties of the closest walls are very different between the zones.

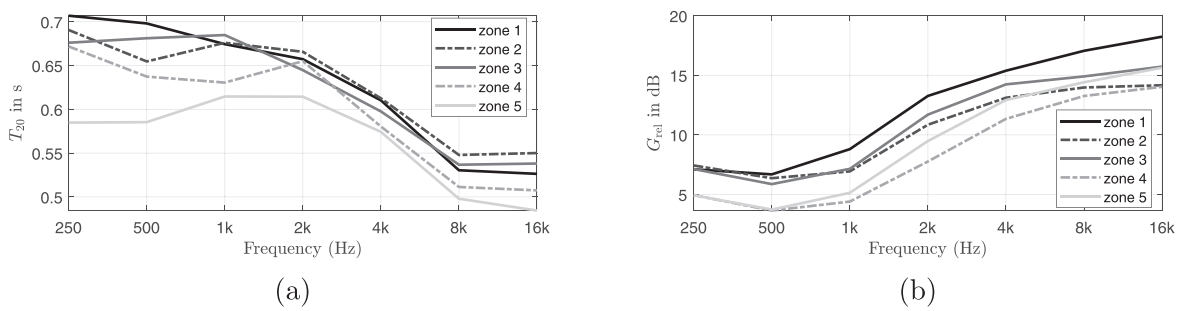

FIG. 2. (a) T_{20} and (b) G_{rel} of the room Arni in octave bands. The indicated zones correspond to those shown in Fig. 1.

TABLE I. Acoustical parameters related to attributes that might aid the zone identification process. All values are at mid frequencies, i.e., averaged over 500 Hz and 1 kHz octave bands.

	Zone				
Parameter	1	2	3	4	5
T20 (s)	0.69	0.67	0.68	0.63	0.60
EDT (s)	0.62	0.62	0.66	0.61	0.61
C50 (dB)	4.56	4.34	2.18	5.45	4.50
G_{rel} (dB)	7.74	6.65	6.51	4.03	4.44
DRR (dB)	−9.38	−7.37	−7.46	−4.49	−5.22

A. Objective acoustic parameters

For future reference and for comparison to similar studies, it is important to quantify the magnitude of differences in acoustic parameters within the room. Such parameters also allow the perceived differences noted in the paired comparison test to be related to physical quantities. For the objective analysis, spatial room impulse response measurements were conducted 1 m behind the loudspeaker using a GRAS VI-50 vector intensity probe. As an example, Fig. 1 shows the placement of the probe in zone 1. The probe consists of six omnidirectional microphones, with one pair on each coordinate axis. An omnidirectional response p was extracted from the topmost microphone. Single-value parameters defined in ISO 3382-1:2009 (2009) were obtained by averaging the 500 Hz and 1 kHz octave band, as recommended in the standard for G, early decay time (EDT), and C_{80}.

First, the reverberation time T_{20} was computed. Figure 2(a) shows the result in octave bands, and Table I shows the single-value reverberation time.[1] Deviations between measurements from different zones are roughly within 10%. Also, the EDT was computed (see Table I).

The direct-to-reverberant ratio (DRR) is given by

$$\mathrm{DRR} = 10\log_{10}\frac{\int_{0\mathrm{ms}}^{5\mathrm{ms}} p^2(t)dt}{\int_{5\mathrm{ms}}^{\infty} p^2(t)dt}. \quad (1)$$

It is inversely correlated with the loudness-related measure G_{rel}, which we defined as

$$G_{\mathrm{rel}} = 10\log_{10}\frac{\int_{0\mathrm{ms}}^{\infty} p^2(t)dt}{\int_{0}^{5\mathrm{ms}} p^2(t)dt}. \quad (2)$$

The range of broadband DRR and EDT can be compared to the range of the room used in Neidhardt's studies (Schneiderwind and Neidhardt, 2019). There, DRR varied between −15.4 and −13 dB when the source was turned away from the measurement microphone. In our room, the DRR value varies slightly more, between −9.4 and −4.5 dB. In the earlier tests, the range of EDT values was 0.42–0.56 s, whereas in the present room, it is only 0.61–0.66 s.

Figure 2(b) shows octave-band values of G_{rel}. Here, zone 4, where the loudspeaker was playing against the absorbing wall, exhibits the lowest values. Zones 2 and 3 show similar values at low frequencies, but zone 2 has slightly less high frequency content, which is expected to induce a subtle coloration difference that participants might notice in the paired comparison test.

Spatial analysis based on short-term time difference of arrival estimates, known as the spatial decomposition method (SDM) (Tervo *et al.*, 2013), presented in Fig. 3 completes the objective analysis. It shows the distinct reflection patterns found at each position and an approximation of the late energy distribution. The measurement from zone 1 exhibits strong early reflections from the left, and zone 2 exhibits them from the right. Zone 3 is more balanced with reflections arriving slightly later from the back. Zone 4 and zone 5 exhibit overall less energy, as seen before. Here, spatial differences become most apparent in the later part of the response, where at zone 4, late energy arrives mostly from the right and in zone 5 from the left, where the walls are most distant.

III. FIRST EXPERIMENT: DIRECT ACOUSTICAL COMPARISON

In the first experiment, a group of listeners was asked to detect whether two renderings came from the same position in a room and provide attributes that supported their decision. In this way, the test not only shows whether the described differences in acoustical attributes between

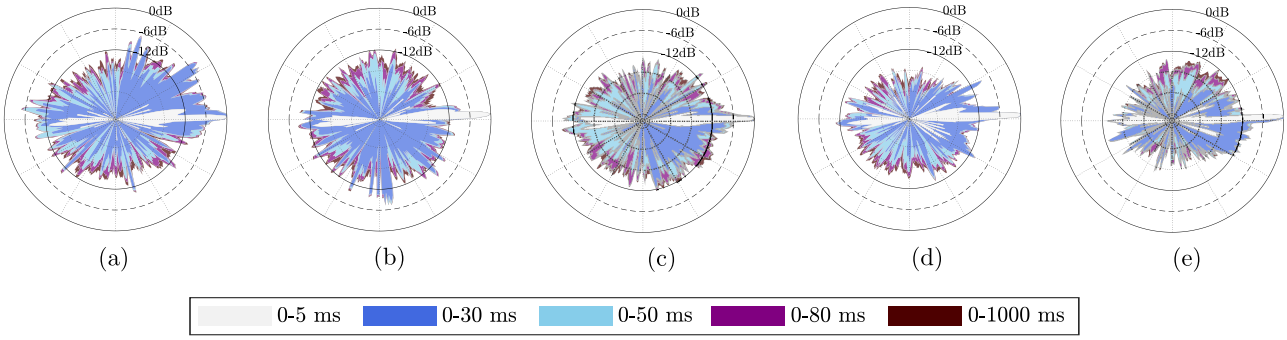

FIG. 3. (Color online) Spatial analysis of the measurements from the five zones. The areas show the directional energy integrated in time windows of increasing length.

https://doi.org/10.1121/10.0013364

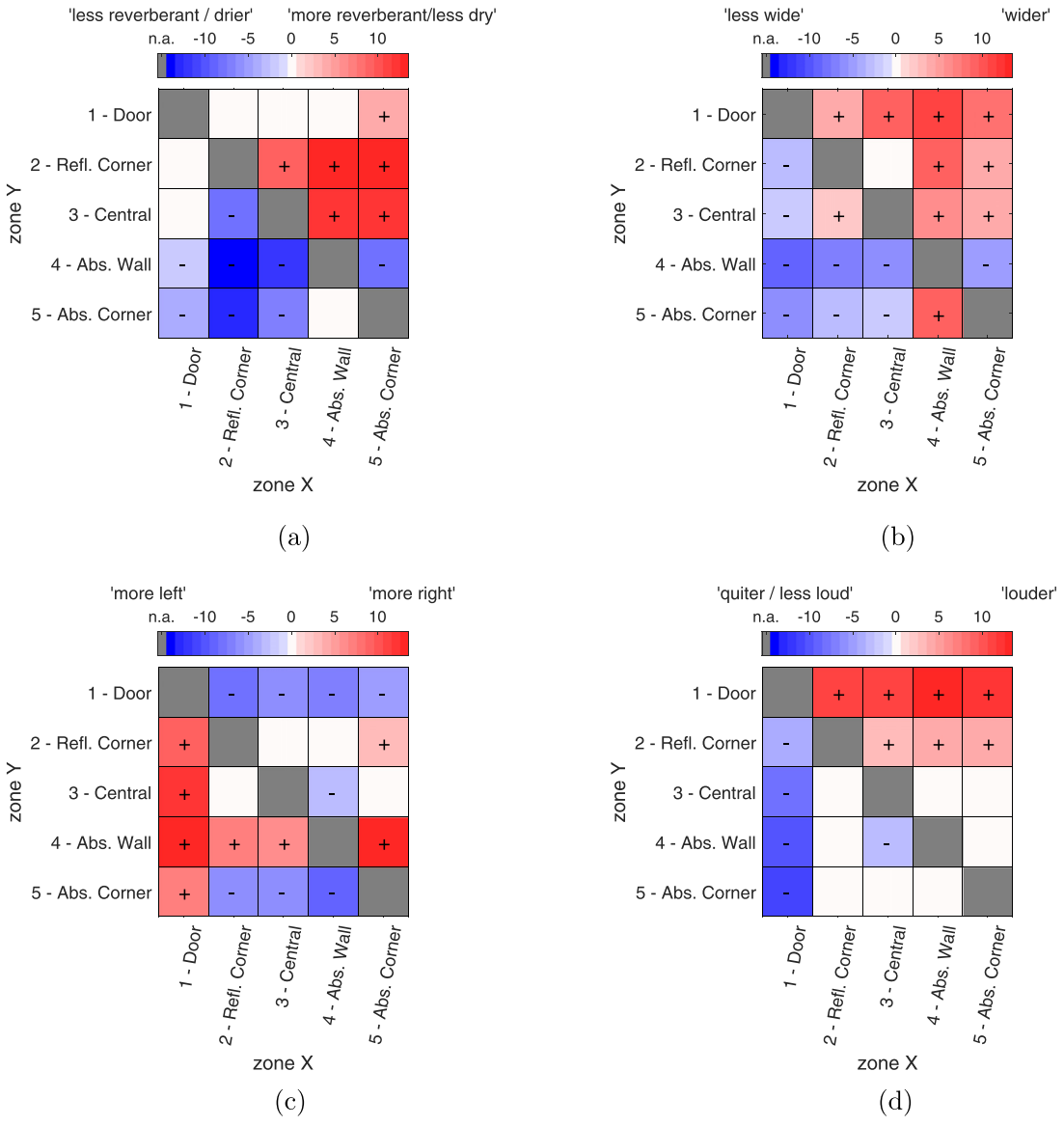

FIG. 4. (Color online) Number of mentions of the four most commonly mentioned attribute pairs. The directed attributes are in reference to X. As an example, a positive entry in (a) means "Y is more reverberant than X."

positions in the room are strong enough to be perceived, but also provides attributes that can later be related to the identification results obtained in the locoscope test. The experiment is a traditional listening test with binaural reproduction of recorded sound environments in which different samples were compared without any visual information.

A. Design

BRIRs were measured with the GRAS KEMAR (Holte, Denmark) head and torso simulator 80 cm behind and 60 cm to the left of the source, facing forward, so that the source appears to the front-right of the listener; see the placement of the dummy head in Fig. 1 for zone 1 as an example. For the experiment, this back-side position was selected because the lower DRR that is found behind a directive source and having lower DRR in one ear than in the other ear could make room acoustical differences easier to perceive, as observed by Schneiderwind and Neidhardt (2019) and Shinn-Cunningham and Ram (2003), respectively. As the direct sound is always the same, a low DRR means that the differences in the reverb are possibly more pronounced.

We wanted to rule out that differences would be detected that are only related to non-ideal positioning and not to the room acoustics, so measurement position and orientation were set up carefully. In addition, prior to convolving the responses with the two signals used in the locoscope experiment, ILDs and ITDs of the direct sound were adjusted to match between the zones 1–5. For this, broadband ILDs were calculated based on the windowed direct sound, and ITDs were determined with sub-sample accuracy using interpolated cross correlation.

The final auralizations were then presented to ten listeners [mean age 28.3 years, standard deviation (SD) = 4.50 years]

in a sound isolated booth using Sennheiser (Wedemark, Germany) HD650 headphones. The test was implemented in MATLAB. The experiment was designed as a triangle test [Zacharov (2018), p. 118] where three auralizations are compared, two of which are the same and one of which is the odd-one-out that is to be selected by the listener. Two different signals were used, one of which was female speech and the other of which was a drum loop. Every pairing of stimuli occurred twice, with each stimulus as the odd-one-out once. This results in 20 trials for the drum signal and 20 trial for the speech signal. In addition to asking which of the three auralizations differed from the remaining two, participants were also asked to provide at least one, and at most four, acoustical attributes they used for making the choice. Examples for such attributes are comparatives such as "louder" or "wider" or short descriptions like "more to the left." The same examples were also given to the participants before the test.

B. Results

Participants were able to find the stimulus that was different from the other two in all but a single trial out of a total of 400 trials (10 participants × 40 trials). This result shows that differences between positions in the room are clearly audible.

Figure 4 shows the number of occurrences for the four most commonly mentioned attributes, which we call "reverberance," "width," "spatial location," and "loudness." When counting them, modifiers were removed, such that a response like "slightly louder" was counted as "louder."

Looking at the results in Fig. 4(a), it becomes clear that zones 1–3 were often perceived as more reverberant than zones 4 and 5. This corresponds well to the differences in reverberation time and DRR shown in Table I. With regard to width, the same pattern is observed, together with a larger width for zone 1, where the loudspeaker faces the reflecting door, than for all the other positions. Also, zone 1 was perceived as louder than all other positions [see Fig. 4(d)]. Accordingly, in Table I, zone 1 exhibits the highest G_{rel}. Last, it is interesting that spatial attributes were mentioned frequently. For example, the sound rendered for zone 1 was perceived to come more from the left than in other zones. This can easily be related to the strong early reflections from the left, seen in Fig. 3(a). Also, a clear spatial difference between zones 4 and 5 was perceived that can also be related to the spatial energy distribution shown in Figs. 3(d)–3(e).

A more detailed table showing responses is found in the supplementary material.[1] Some of the attributes are similar to those found in the study about different positions in a concert space (Băcilă and Lee, 2021). There, loudness, width, level of reverb, reverb direction, distance, and brightness of the reverb were amongst the most salient differences between positions. However, in their study, attributes were further differentiated, for example, in "source width," "reverb width," and "envelopment." Some attributes only apply to larger spaces, like "echo brightness" and "echo direction."

IV. SECOND EXPERIMENT: LOCOSCOPE TEST

The locoscope design is based on a three-dimensional (3D) model of an existing room. The model was created by means of 3D scanning. For this, the iOS application *3D Scanner App*[2] was employed, which uses the LiDAR depth camera system available in Apple (Cupertino, CA) iPad Pro models (2020). After scanning, the model was imported into Unity as an .fbx file. Since during each trial, only one of five loudspeakers and one of five chairs was visible, a loudspeaker and a chair model were created separately, such that they could separately be displayed or not.

During the test, the virtual room model needed to be aligned to the real room. For this, the experimenter performed an alignment routine by touching four pre-defined points in the real room using one of the controllers before the start of the experiment. After acquiring the positions of these points, Procrustes analysis was applied to rotate and shift the virtual room to match the real one. Figure 5 shows two views into the real room and the aligned virtual room.

To run the test, a HMD (Quest 2, meta, Menlo Park, CA) was placed on the participant's head, showing the room model. A silicone cover ensured that the device was optically sealed, which was of critical importance. The acoustical influence of the Quest 2 itself was expected to be small. Separate experiments have shown that merely a small amount of coloration was perceived for sound from mostly frontal directions, and localization was hardly affected (Ahrens *et al.*, 2019; Lladó *et al.*, 2022). Such minor impairments are assumed to be negligible for the given task in which the listener can move in 6DoF. Note that the internal speakers of the device are never used in the test.

With the HMD in place, participants were asked to sit down on a chair that is attached to a platform with large rubber wheels. The HMD was set to show a black screen, and participants were randomly moved and spun around in the room, until they lost orientation. Using the chair, such confusion was typically achieved after a few turns. Then, the chair was placed in one of the five loudspeaker zones shown in Fig. 1. Once the listener was placed in one of the zones, the experimenter pushed a button that triggered two events: First, the loudspeaker in the zone in which the participant had been placed was activated. Second, the HMD then showed the virtual room, and the trial began. Crucially, the zone they found themselves in visually did not necessarily match the zone they were physically placed in. The participants could now walk around the loudspeaker in the zone and use the joystick on the handheld controller to switch their position in the visual model that was shown in the HMD. In this way, they effectively teleported between the five different zones in the virtual world. When doing so, the loudspeaker model remained at a constant position relative to the listener; only the surrounding world was rotated and shifted. It was important that the position relative to the loudspeaker stayed the same while switching between zones, so that the direct sound location did not provide any cue, except in two control trials described below. The boundaries of the zones

https://doi.org/10.1121/10.0013364

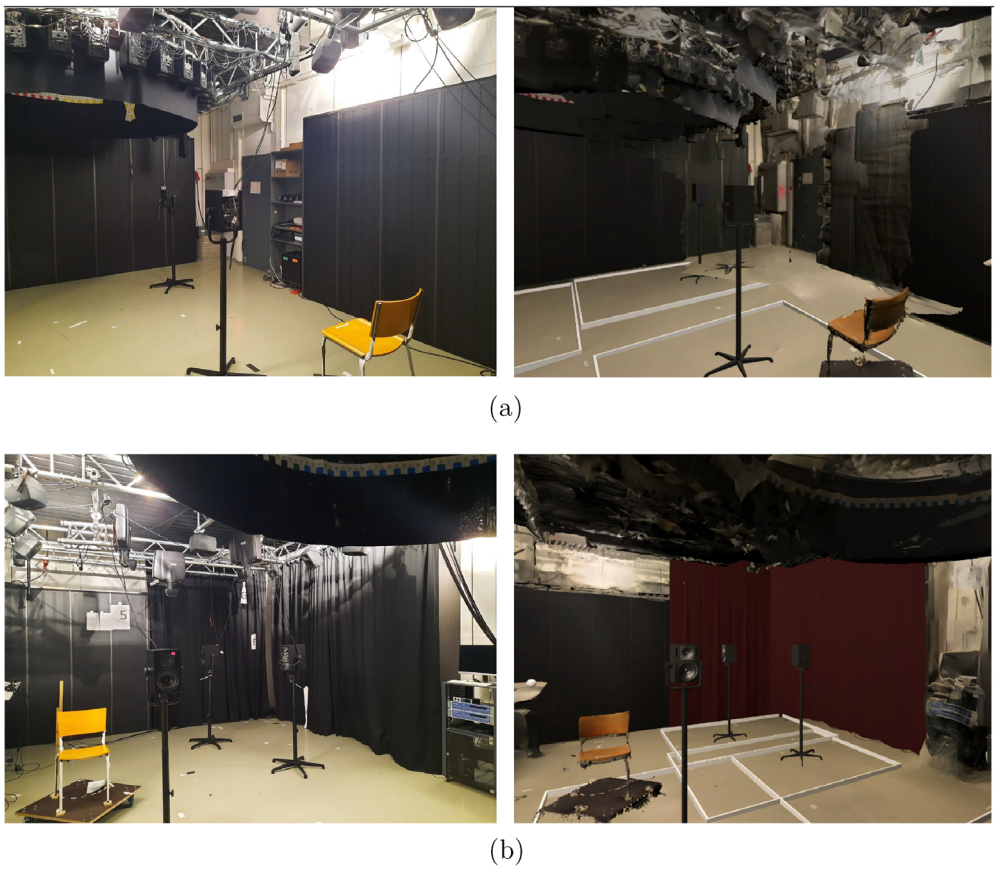

FIG. 5. (Color online) Pictures of the real room on the left and screenshots of the virtual room model on the right.

were marked in the virtual model, and participants were instructed not to leave them.

Participants were then asked to select the zone (of the five different options) that corresponds to what they believe was their position in the real room, based on the sound they heard. After making the selection, they sat back down on the chair. Then, the screen turned black, and the listener was moved into another zone for the next trial. Again, a random path was followed, while slowly spinning the chair, in order for the participant to lose orientation. A video showing a participant during one trial can be found online.[3]

Upon the start of a test, the task was explained to the participants while seated in the center of the room, at zone 3. In the test, there was one trial in each of the five zones with the loudspeaker in the zone reproducing an anechoic female speech recording ($L_{\mathrm{Aeq}} = 65$ dB at zone 3, 1 m on-axis) and one trial each where a drum loop was used ($L_{\mathrm{Aeq}} = 72$ dB at zone 3, 1 m on-axis). For calibration, all loudspeakers were moved to zone 3, such that after placing them back in their designated position, all occurring level differences were only of room acoustical origin. As control conditions, all zones occurred once in silence. The intention was to see if background noise or non-acoustical factors such as smell would permit self-localization. Moreover, during two trials, a loudspeaker on the desk in the back right corner was activated instead of the loudspeaker in the zone (see "Desk" in Fig. 1). During these two special control trials, which always took place in zones 1 and 2, participants could switch between visuals in the same way as before, but they heard the more distant sound source outside of the walking zone. In contrast to the other trials, where the loudspeaker always remained in the center of the zone, it was possible to match the direct sound location to the location of the loudspeaker in the visuals in these trials. This allows for checking the influence of a localizable source that appears at different locations between the positions presented visually in the lococope.

With five trials per signal condition, five for the silent condition, and the two trials with the loudspeaker on the desk, there were 17 trials per participant, presented in a random order. Most participants examined the scenes closely, and completing the experiment itself took between 30 and 60 min, after which most participants felt exhausted. After half the trials, a break was offered, for which participants were moved to a neutral position in the center of the room. Some participants needed extra breaks, and three participants completed the test in two sessions.

Twenty participants with a mean age of 27.8 years (SD = 4.6) took part in the experiment. They were all staff and students associated with the Aalto Acoustics Lab, but care was taken to include participants with different experience levels. Listeners were taken from the same pool as in the experiment described above. Before the test, participants

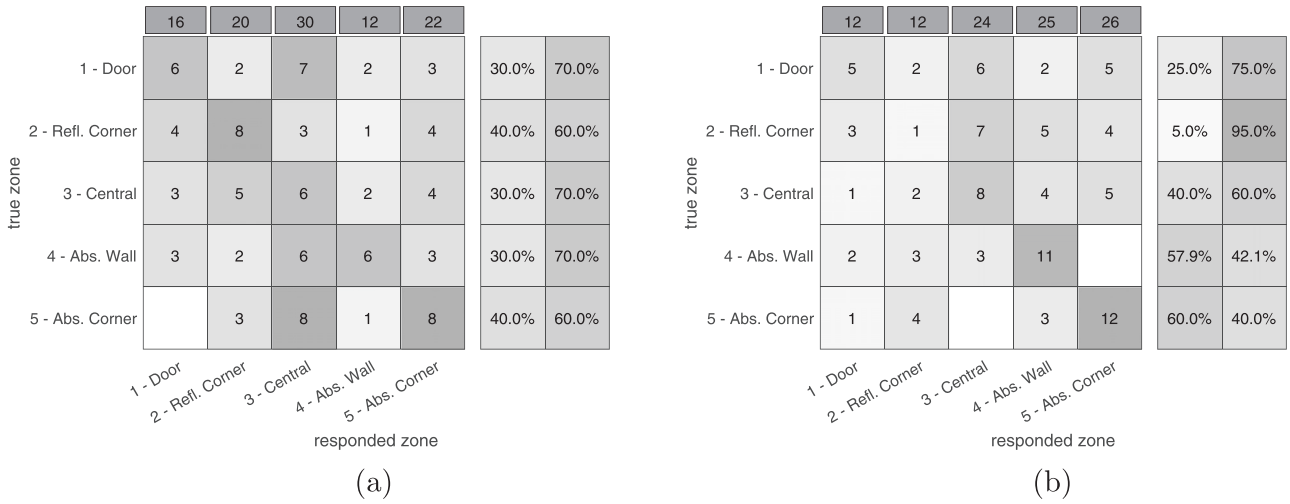

FIG. 6. Confusion matrices showing the true, physical positions and the participants' responses for (a) speech and (b) drums. The right column shows the percentage of correct responses per true physical position. The position in the absorbing corner (zone 5) was recognized the most. The two positions in which the loudspeaker was facing toward the center of the room (zone 2 and zone 3) were commonly confused with each other.

signed an informed consent form including information about the purpose of the test, publication of the data, the prospect of experiencing disorientation, and the option to leave the test at any time. After the main experiment, a questionnaire was administered. It included questions about the experience during the test, as well as personal background in acoustics, audio engineering, and professional musical training.

V. RESULTS

A. Subjective experiences with the design

In the design, it is instrumental for the visual rendering to be capable of creating a PI, i.e., the sense of being at a certain place in the room, at least in cases where participants think they are in the correct zone. Therefore, participants were asked "How strongly was the virtual model capable of creating the illusion of being at a different place in the room?" taking the complete test into account. The mean response on a scale from 1 to 10 was 8.37 [SD = 1.71, median (Mdn) = 9], indicating that it was often possible to evoke PIs.

When asking participants about the difficulty of the test, the mean score was 7.8 (SD = 1.44, Mdn = 8), which shows that most participants found the test difficult. Furthermore, when asked how the experience felt as a whole, the design received positive feedback ("cool," "new," "a lot of fun," "fairly seamless," "quite immersive," "interesting"). However, some participants also pointed out that it was "challenging" and "disorienting," possibly indicating that they did not feel completely comfortable in the virtual environment, which could potentially influence the performance.

B. Identification performance

Turning toward the identification results, it appears that listeners were able to identify only specific positions relatively well, while most positions were nearly impossible to identify (see Fig. 6). The overall proportion of correct answers in the speech and the drum trial together was $p_c = 0.36$ (SD = 0.18), compared to the guessing probability of $p_{guess} = 0.2$.

Most notably, subjects were often able to recognize the positions close to the absorbing corner in the case of the drum signal; zone 5 had a proportion of correct answers of $p_c = 0.6$, and in zone 4, the score was $p_c = 0.58$. A binomial test allows for comparing this identification rate to the probability of obtaining at least this number of correct responses purely by guessing. In case of $p_c = 0.6$ and $p_c = 0.58$ for 20 decisions between five alternatives, this probability is $p < 0.001$, which makes it unlikely that the result has been obtained by chance.

With $p_c = 0.4$, the central position was recognized second most often, with probability to obtain the result from guessing at $p = 0.032$. For the drum signal, the position next to the reflective corner (zone 1) was almost never recognized; instead, it was commonly confused with the central position in the room (zone 3).

Overall, there was no significant difference between the drum signal ($p_c = 0.37$) and the speech signal ($p_c = 0.34$), but for the speech signal, there were fewer differences between the positions. Zone 2 and zone 5 were recognized with $p_c = 0.4$. All other positions were recognized with $p_c = 0.3$ cases, which may be obtained by pure guessing with a probability of $p = 0.37$. For both signals, zone 4 and zone 5 were rarely confused with each other. Furthermore, some asymmetries can be noted. For example, zone 5 was mistaken for zone 1 only once in the whole experiment; however, vice versa, zone 1 was thought to be zone 5 eight times in total.

Apart from the signal conditions, it needs to be mentioned that participants performed above chance for the silent condition ($p_c = 0.3$). The performance in the silent condition can partly be explained by participants trying to localize faint background noises due to some electrical

https://doi.org/10.1121/10.0013364

installations or the radiators in the room. However, it is not to be said that these noises were always assigned to the correct source, and even those participants who claimed to have listened for noise in the silent control condition reported that it was inaudible during signal conditions. Another option is that the sound of footsteps or the noise of moving the chair around was used as a cue. It is unlikely that participants have used any non-acoustical cues, like smell or remaining vision through the sides of the sealed headset, as none of the participants mentioned any in the questionnaire. Also, before the test, all participants formally agreed to immediately disclose if non-acoustical cues would reveal their true position to them.

In the second control condition, where the loudspeaker on the desk was active, all participants were always able to recognize their position with great ease ($p_c = 1.0$).

C. Listener experience

As in earlier experiments (Klein *et al.*, 2017), inter-individual performance differences were large. While the best participant achieved $p_c = 0.77$, there was also a participant with no correct answers at all (except in the control condition with the loudspeaker on the desk). A Shapiro–Wilk test indicates that the percentage of correct answers may be normally distributed amongst the participants.

It is vital for the evaluation of future VR applications to establish the relationship between performance differences and the listeners' experience. In general, assessing overall listener experience is a difficult matter in itself. Recently, von Berg *et al.* (2021) have conducted a study in which participants with varying experience in music, room acoustics, and audio engineering performed a perceptual task and a room recognition task. It was shown that formal music education, experience with music recording, and academic acoustic knowledge correlate with the performance in both tasks in cases where reverberation time was varied between trials, but less so when the spectral envelope of the reverberation was changed.

In our study, we asked for the number of years of experience in the same three fields (professional music training, professional audio engineering, academic acoustics). Contrary to what was first hypothesised after conducting the test with eight participants, a multiple regression analysis based on those three factors did not yield any significant correlations; the adjusted R^2 indicated that only roughly 1% of the outcome may be explained by these factors.

The failure of multiple linear regression shows that a linear relationship between years of experience in the three field fields and performance in the test is not a reasonable assumption. However, looking at the answers concerning experience, it was easily possible to distribute the participants into two groups: one group with participants that had less than 2 years of experience in each respective field and one that had 2 or more years.

A two-sample t-test indicates that the difference in performance between participants with less than 2 years of experience in acoustics research (with a mean proportion correct of $p_c = 0.23$) and the more experienced listeners ($p_c = 0.41$) is relatively unlikely to have emerged from limited sampling, $t(18) = 2.18$, $p = 0.043$. When partitioning the participants according to audio engineering experience ($p_c = 0.29$ with and $p_c = 0.43$ without), a difference exists, but the limited number of participants permits no strong conclusions, $t(18) = 1.76$, $p = 0.096$. No conclusions are possible about musical training ($p_c = 0.33$ for non-experienced and $p_c = 0.41$), $t(18) = 0.92$, $p = 0.37$.

These results indicate that there is a tendency for listeners with some listening experience to perform better at the task, especially regarding experience in acoustics research. However, note that not all experienced participants performed well, so the relationship is far from causal.

The experiment was designed as such to avoid habituation throughout the experiment, which is why every position/stimulus combination occurred only once and the order was fully randomized. No strong habituation effect was found; linear regression between the position of the stimulus in the test and the percentage of correct answers resulted in $R^2 = 0.023$, $p = 0.303$.

VI. DISCUSSION AND IMPLICATIONS

With the results of the locoscope test available, it is possible to refer back to the objective analysis and the attributes collected in the direct comparison test. Specifically, we highlight particular cases that were confused most or least often.

A good example is zone 1, which was always perceived to be louder than all other positions in direct comparison [see Fig. 4(d)], indicated by the highest G_{rel} amongst the samples. Nevertheless, zone 1 was commonly confused with all the other positions in the locoscope test. This shows that the high loudness caused by the early reflections from the front and from the side was not amongst the expectations that participants were able to form from the visual input. The same is true for the more subtle spectral differences between zone 2 and zone 3. The recognition of absolute loudness or coloration is not possible without direct comparison, but it might increase after many training trials, which was deliberately avoided in the locoscope test.

An important consideration is that zone 1, in comparison to zone 2 and zone 3, was not only louder, but also the pattern of early reflections strongly differed between the three positions [see Figs. 3(a)–3(c)], which has led participants to perceive the sound at zone 1 more from the left than in zone 3 in direct comparison [see Fig. 4(c)]. Regardless of this strong spatial difference in early reflections, zone 1 was often confused with zone 3 in the locoscope test.

Figure 4(a) shows that zone 4 and zone 5 were described as less reverberant or more dry than the other positions, which some participants were apparently able to associate with the presence of the heavy curtains. For both drum and speech stimuli, zone 4 and zone 5 were rarely confused with each other. One cue for this might have been the

DRR. Experienced participants may have concluded that the room is excited less when a directional source like a studio loudspeaker is playing directly at an absorbing wall. Another explanation is that spatial cues were the most revealing. The spatial analysis shows that at zone 5, there was much more late energy from the left side [see Fig. 3(e)], which opened toward the room, than at zone 4, where the walls on the left side were absorptive.

It is interesting that zone 4 and zone 5 were recognized much more reliably with the transient drum signal than with the speech signal. A possible explanation is that in the transient case, it is easier to focus on the directional properties of the reverberation alone, as the directional energy is not integrated so strongly over time as with a more continuous signal. The effect of the signal was not very strong for the other positions (the overall worst performance was seen for zone 2 and the drum signal), so that even though it can be expected that the echo threshold is lower for the more transient signal, early reflections were not easily recognized.

All in all, the results show that for a group of listeners not specifically trained in echolocation, identifying one's own position in a room based on clearly audible room acoustic cues can be very difficult, even when moving freely in that room. Differences in the particular early reflection patterns confirmed by spatial analysis could not be used to assert the position in the room.

Strictly speaking, the presented results are only valid for the case of this particular room, which is a limitation of the presented study. The tested room provided relatively large position dependent differences when compared to many rooms, but it is possible that spaces exist in which even larger differences would make the audiovisual association task easier. However, based on the present results, it appears unlikely that the perfect performance observed in the direct comparison task would be found when performing the lococope task in another room. Nevertheless, the study should be replicated in different rooms. Furthermore, it should be noted that the experiment focused on passive self-localization. As future VR applications might also include auralizations of self-produced sound, conducting a study in which self-produced sound is allowed could give relevant insights.

Future research could also focus on studying the underlying cognitive principles when comparing the perceived sound with the expectation of a listener about how the source would sound given the visual information. Models for auditory cognition and auditory memory can serve as a starting point. A common model (Ashcraft and Radvansky, 2014) assumes the existence of auditory sensory memory, in which all incoming sound is briefly stored before it is encoded into verbal categories in working memory, which in turn interfaces with long-term memory. In our experiments, the good discriminability in direct comparison would rely on the early stage auditory sensory memory, while the lococope test relies on encoding information about the heard sound and comparing it with information from the visual system and acoustic knowledge from long-term memory.

Note that in contrast to previous studies, participants did not receive prior training in our experiments. Furthermore, by fully randomizing positions and stimuli, and by the fact that each combination only occurred once, the effect of habituation during the experiment was successfully kept low. The fact that in other tests participants tended to perform better after extensive training, together with the observation that many perceivable differences exist within a room like the tested one, might in fact be explained by the mentioned cognitive model: During training, a listener might learn to attend to acoustical features more strongly and thereby enable their encoding into working memory. It is an open question whether auditory memory and cognitive processing for rooms only depends on verbal categories or if, as for pitch (Deutsch and Deutsch, 1975), non-verbal memory for room acoustics exists.

It would be possible to use the presented design to actively explore the effects of learning, for example, by providing a training phase, where feedback is provided, even though feedback about correct rendering would not be provided in VR applications either.

Returning to specific results and implications for VR, identification was best when using a loud, transient drum signal at two positions that had lower DRR than the others and had a distinct directional distribution of the late reverb (zone 4 and zone 5). Another position, with a strong lateral early reflection and higher loudness, was not recognized as often (zone 1). This observation could lead to the new hypothesis that listening to the reverberant tail as its own auditory event, following a transient excitation, with its own spatial distribution, is more important for self-localization by untrained listeners than the effects of early reflections. It is likely that the comb-filtering caused by such reflections is hard to perceive in a real room and that the image shifts that were reported in the paired comparison test are easily overwritten by vision in the form of a room acoustical ventriloquism effect. This hypothesis should be confirmed in future tests, for instance, using tracked binaural rendering together with the lococope application. For virtual acoustics in VR, confirmation could mean that incorporating some directional dependency in a unique global reverb, for example, creating direction dependent damping in the vicinity of absorbing walls or open windows, might be more important than simulating early reflections correctly.

Another interesting observation was that when the speaker on the desk was playing, the task was trivial for all participants. This shows that self-localization is much more easily achieved by localizable external sound sources than by attending to room acoustics. Note that when Picinali *et al.* (2014) studied the exploration of virtual environments by blind people, localizable sound sources were present in the simulation as well. It is an open question whether an increase in self-localization ability achieved by adding external sound sources might also increase the chance to create PIs in VR.

VII. CONCLUSION

We have introduced an experimental design using a VR application that we call locoscope. By letting listeners physically walk around, examining the acoustics in zones of a real room, the design is more ecologically valid than similar experiments where the sound is rendered virtually over headphones and the participants are not able to physically move around. We have also shown the important difference between two paradigms when evaluating virtual acoustics for VR. On the one hand, room acoustical differences between several positions in a room were easily audible in direct comparison. The attributes used to describe these differences were mostly related to loudness, localization, width, coloration, and reverberance. On the other hand, in the locoscope test, participants were rarely able to relate the distinct acoustic features of positions in the room to the same positions presented visually. Only some listeners were able to recognize some specific positions, mostly with a loud and transient drum sample.

In other words, on many occasions, participants were, for example, standing directly next to a wall, thinking they were in the middle of the room, despite listening carefully with their own ears. This leads us to answer the introductory question "Can you hear where you are in a room?" with "no" for many participants in the tested room.

Furthermore, our discussion highlights that self-localization based on acoustics is promoted by acoustic experience, where participants with acoustic experience tend to perform better. Also, we have seen that the ability to perform self-localization based on localizable sounds in the room is much higher than when using room acoustic cues only. While it is clear that the direction of a direct sound should be rendered correctly in VR applications, the results suggest that cases exist in which the importance of position dependent rendering of the acoustics for a general audience is limited. It should be noted that the present results concern non-interactive sound sources in the proximity of a listener. Also, it is always possible that other perceptual or cognitive effects of incorrect rendering will be found in future research.

ACKNOWLEDGMENTS

Many thanks to Janne Pietilä for the first implementation and for running the pilot test and to Juuso Tolonen for building the moving platform. This research has received funding from the European Union's Horizon 2020 research and innovation program under Marie Skłodowska-Curie Grant No. 812719.

[1]See supplementary material at https://www.scitation.org/doi/suppl/10.1121/10.0013364 for all parameters in octave bands, the stimuli used in the direct comparison test, and a more detailed table of mentioned attributes.
[2]The scanning application used for making the model is available at https://www.3dscannerapp.com/ (Last viewed July 28, 2022).
[3]Audio and video examples are available at http://research.spa.aalto.fi/publications/papers/jasa_whereyouare/ (Last viewed July 28, 2022).

Ahrens, A., Lund, K. D., Marschall, M., and Dau, T. (**2019**). "Sound source localization with varying amount of visual information in virtual reality," PLoS ONE **14**(3), e0214603.
Ashcraft, M. H., and Radvansky, G. A. (**2014**). *Cognition*, 6th ed. (Pearson Education, Boston).
Băcilă, B. I., and Lee, H. (**2021**). "Listener-position and orientation dependency of auditory perception in an enclosed space: Elicitation of salient attributes," Appl. Sci. **11**(4), 1570.
Deutsch, D., and Deutsch, J. A. (**1975**). "The organization of short-term memory for a single acoustic attribute," in *Short Term Memory* (Academic, New York), pp. 107–152.
Engel, I., Henry, C., Amengual Garí, S. V., Robinson, P. W., and Picinali, L. (**2021**). "Perceptual implications of different Ambisonics-based methods for binaural reverberation," J. Acoust. Soc. Am. **149**(2), 895–910.
ISO 3382-1:2009 (**2009**). "Acoustics-measurement of room acoustic parameters—I: Performance spaces" (International Organization for Standardization, Geneva, Switzerland).
Klein, F., Neidhardt, A., Seipel, M., and Sporer, T. (**2017**). "Training on the acoustical identification of the listening position in a virtual environment," in *Proceedings of the 143rd AES Convention*, October 18–21, New York.
Kolarik, A. J., Cirstea, S., Pardhan, S., and Moore, B. C. (**2014**). "A summary of research investigating echolocation abilities of blind and sighted humans," Hear. Res. **310**, 60–68.
Litovsky, R. Y., Colburn, H. S., Yost, W. A., and Guzman, S. J. (**1999**). "The precedence effect," J. Acoust. Soc. Am. **106**(4), 1633–1654.
Lladó, P., McKenzie, T., Meyer-Kahlen, N., and Schlecht, S. J. (**2022**). "Predicting perceptual transparency of head-worn devices," J. Audio Eng. Soc. **70**, 585–600.
Lübeck, T., Arend, J. M., and Pörschmann, C. (**2022**). "Binaural reproduction of dummy head and spherical microphone array data—A perceptual study on the minimum required spatial resolution," J. Acoust. Soc. Am. **151**(1), 467–483.
Neidhardt, A. (**2016**). "Perception of the reverberation captured in a real room, depending on position and direction," in *Proceedings of the 22nd International Conference on Acoustics*, September 5–9, Buenos Aires, Argentina.
Neidhardt, A., Fiedler, B., and Heinl, T. (**2016**). "Auditory perception of the listening position in virtual rooms using static and dynamic binaural synthesis," in *Proceedings of the 140th Audio Engineering Society Convention*, June 4–7, Paris, France.
Paasonen, J., Karapetyan, A., and Plogsties, J. (**2017**). "Proximity of surfaces—Acoustic and perceptual effects," J. Audio Eng. Soc. **65**(12), 997–1004.
Picinali, L., Afonso, A., Denis, M., and Katz, B. F. G. (**2014**). "Exploration of architectural spaces by blind people using auditory virtual reality for the construction of spatial knowledge," Int. J. Hum. Comput. Stud. **72**(4), 393–407.
Prawda, K., Schlecht, S. J., and Välimäki, V. (**2020**). "Evaluation of reverberation time models with variable acoustics," in *Proceedings of the 17th Sound and Music Computing Conference*, June 24–26, Torino, Italy.
Schneiderwind, C., and Neidhardt, A. (**2019**). "Perceptual differences of position dependent room acoustics in a small conference room," in *Proceedings of the International Symposium on Room Acoustics*, September 15–17, Amsterdam, Netherlands, pp. 499–506.
Shinn-Cunningham, B., and Ram, S. (**2003**). "Identifying where you are in a room: Sensitivity to room acoustics," in *International Conference on Auditory Displays*, pp. 21–24.
Slater, M. (**2009**). "Place illusion and plausibility can lead to realistic behaviour in immersive virtual environments," Philos. Trans. R. Soc. B **364**(1535), 3549–3557.
Tervo, S., Pätynen, J., Kuusinen, A., and Lokki, T. (**2013**). "Spatial decomposition method for room impulse responses," J. Audio Eng. Soc. **61**(1), 17–28.
von Berg, M., Steffens, J., Weinzierl, S., and Müllensiefen, D. (**2021**). "Assessing room acoustic listening expertise," J. Acoust. Soc. Am. **150**(4), 2539–2548.
Zacharov, N. (**2018**). *Sensory Evaluation of Sound*, 1st ed. (CRC, Boca Raton, FL).