acoustics

MDPI

Article

Real and Virtual Lecture Rooms: Validation of a Virtual Reality System for the Perceptual Assessment of Room Acoustical Quality

Angela Guastamacchia [1,*], Riccardo Giovanni Rosso [2], Giuseppina Emma Puglisi [1], Fabrizio Riente [3], Louena Shtrepi [1] and Arianna Astolfi [1]

[1] Energy Department, Politecnico di Torino, 10129 Torino, Italy; giuseppina.puglisi@polito.it (G.E.P.); louena.shtrepi@polito.it (L.S.); arianna.astolfi@polito.it (A.A.)
[2] Computer Science Department, Università degli Studi di Torino, 10124 Torino, Italy; riccardo.rosso718@edu.unito.it
[3] Department of Electronics and Telecommunications, Politecnico di Torino, 10129 Torino, Italy; fabrizio.riente@polito.it
* Correspondence: angela.guastamacchia@polito.it

Abstract: Enhancing the acoustical quality in learning environments is necessary, especially for hearing aid (HA) users. When in-field evaluations cannot be performed, virtual reality (VR) can be adopted for acoustical quality assessments of existing and new buildings, contributing to the acquisition of subjective impressions in lab settings. To ensure an accurate spatial reproduction of the sound field in VR for HA users, multi-speaker-based systems can be employed to auralize a given environment. However, most systems require a lot of effort due to cost, size, and construction. This work deals with the validation of a VR-system based on a 16-speaker-array synced with a VR headset, arranged to be easily replicated in small non-anechoic spaces and suitable for HA users. Both objective and subjective validations are performed against a real university lecture room of 800 m^3 and with 2.3 s of reverberation time at mid-frequencies. Comparisons of binaural and monoaural room acoustic parameters are performed between measurements in the real lecture room and its lab reproduction. To validate the audiovisual experience, 32 normal-hearing subjects were administered the Igroup Presence Questionnaire (IPQ) on the overall sense of perceived presence. The outcomes confirm that the system is a promising and feasible tool to predict the perceived acoustical quality of a room.

Keywords: virtual reality system; 3rd-order ambisonics; reverberant educational spaces; perceived acoustical quality; sense of presence; HAs; audiovisual scene; acoustical validation

check for updates

Citation: Guastamacchia, A.; Rosso, R.G.; Puglisi, G.E.; Riente, F.; Shtrepi, L.; Astolfi, A. Real and Virtual Lecture Rooms: Validation of a Virtual Reality System for the Perceptual Assessment of Room Acoustical Quality. *Acoustics* **2024**, *6*, 933–965. https://doi.org/10.3390/acoustics6040052

Academic Editors: Jian Kang, Dadi Zhang and Massimiliano Masullo

Received: 13 September 2024
Revised: 20 October 2024
Accepted: 25 October 2024
Published: 30 October 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The acoustical quality of educational environments is crucial for creating comfortable and effective learning spaces. Poor acoustics can negatively impact comfort and disrupt the learning process by making it difficult for listeners to focus and engage [1]. Additionally, cognitive aspects are closely linked to sound; excessive noise or poor speech clarity can increase cognitive load, thereby increasing the mental effort required to process auditory information [2]. Thus, a key goal in acoustic design should be to achieve high speech intelligibility [3], which is especially vital for vulnerable populations such as hearing-impaired (HI) individuals and hearing aid (HA) users [4].

Hence, to ensure inclusive and supportive learning spaces for all listeners, techniques that assess the perceived speech intelligibility should be employed during both the design of educational environments and the acoustic evaluation of existing spaces. Additionally, understanding how occupants' acoustic perceptions vary within the same environment and identifying the factors that cause these differences is crucial for designing human-centered spaces that promote overall well-being.

Recent advancements in virtual reality (VR) techniques offer exciting new possibilities for studying subjective perception, particularly in the realm of auditory experiences, mimicking everyday listening payloads involving multiple sound sources from different directions in possible reverberant settings [5]. The development of increasingly optimized methods for creating spatialized virtual acoustic environments (VAEs) enables researchers to effectively investigate how the acoustics of different environments influence occupants' well-being. Subjective tests can be conducted in real time within controlled laboratory settings, allowing researchers to examine how minor modifications to a space can impact human perception. Furthermore, VR provides a powerful tool for the subjective assessment of acoustics in both existing and planned environments.

1.1. State of the Art: VR Systems

Currently, various technologies and spatial audio reproduction methods exist for rendering a 3D auditory scene at the listener's ears, each entailing different levels of hardware, processing complexity, and costs, along with their own advantages and disadvantages.

Binaural technology relies on simple headphones or two-speaker arrays with cross-talk cancellation (CTC) filters and the usage of head trackers to recreate a physically correct sound field responsive to listener's head movements [6]. However, it demands significant real-time computational efforts and nearly impractical conditions, such as anechoic listening rooms and personalized head-related transfer function (HRTF) measurements, to mitigate underlying drawbacks like front–back confusions, externalization, and coloration issues [6–10]. In contrast, rendering methods such as wave field synthesis (WFS) [11], nearest-speaker panning (NSP) [12], vector base amplitude panning (VBAP) [13], distance-based amplitude panning (DBAP) [14], multiple-direction amplitude panning (MDAP) [15], and high-order ambisonics (HOA) [16] aim to reconstruct the sound field either physically or perceptually by employing various multi-speaker array layouts. These approaches accommodate listener movements without requiring head trackers or customized HRTFs, leading to strongly reduced real-time processing costs.

The WFS is designed to reproduce physically authentic sound fields, which are particularly recommended in the case of applications that involve listeners with HAs [8,12]. However, while WFS VAE reproduction covers an extended area, suitable for a multi-listener audience up to a maximum aliasing frequency [6], it requires a larger number of speakers compared to other techniques, such as VBAP and HOA, to achieve a similar number of artifacts [12], leading to higher costs and greater implementation complexity. Furthermore, the correct installation of speakers for WFS requires a precise and uniform arrangement around the listening area [11], which can be challenging to achieve in pre-existing or spatially constrained environments. Nevertheless, all multi-speaker-based spatial audio reproduction methods share a common limitation in their spatial resolution, which is constrained by the finite number of speakers used [12]. However, despite this shared limitation, all panning rendering methods, focusing on replicating the perceptual attributes of an acoustical scenario, differ in the number of speakers and spatial layouts required to achieve comparable performance in terms of sweet spot extension (i.e., the region within which the sound field is accurately replicated—which can be very small for some configurations, limiting the allowed listener's movements), sound source width and localization, reproduction of moving sources, and other spectral artifacts.

The NSP is the simplest rendering technique that pans the virtual source to the speaker with the minimum angular distance to it. Thus, for each sound source, a single speaker is activated, which minimizes spectral artifacts [17] and makes NSP free of spatial aliasing for single sources [12]. However, the localization error is strongly determined by the spatial density of the speakers [17] and NSP falls short in replicating diffused and reverberant sound environments and may still lead to spatial aliasing artifacts when multiple sound sources are reproduced [12].

With VBAP, a maximum of three speakers are activated simultaneously to render a single source. In this way, VBAP is more robust against localization errors at high

frequencies and in the case of listener's off-center from the sweet spot than other techniques like HOA, especially when HAs are used [18]. However, VBAP suffers from coloration effects and unsteady source width that depend on the virtual panning position in relation to the speaker arrangement [19,20], as well as dynamic sound reproduction issues [15]. Additionally, it requires careful consideration prior to installation, as it is highly dependent on a fixed and optimal speaker setup [17].

As an extension of the VBAP, the MDAP is thought to mitigate the spectral artifacts associated with VBAP by reproducing a virtual source as superimposition of the VBAP results for B panning directions uniformly distributed around the desired virtual source location [15], which helps to reduce localization error in the sweet spot. In case of HA wearers, it should be preferred to VBAP when dynamic sounds and not-off-center listeners are involved [18].

For applications that require a wide listening area, DBAP can be a valuable tool. It reproduces a virtual source by assigning to the speakers different gains that are inversely proportional to the speaker distance from the desired virtual source location [8]. As a result, its rendering does not depend on any specific listening position, thus no specific constraints are required for the speakers' layout [14]. Compared to VBAP, it results in similar perceived localization errors but fewer discontinuity issues when dealing with moving sounds [15].

Finally, HOA relies on a completely different paradigm for rendering VAEs, based on the sound field decomposition into spherical harmonics [21], which results in almost opposite pros and cons. The internal representation of the sound field is speaker-independent and can be properly decoded using different techniques for any kind of speaker layout [22], making this technique very versatile in terms of space and cost requirements. The level of reproduction accuracy and sweet spot extension, in terms of space and maximum cut-off frequency, are related to the ambisonic truncation order (M), which in turn determines the number of speakers (N) needed, as defined by the formula: $N = (M+1)^2$ [21]. In HOA, all speakers are driven simultaneously, with different gains derived from the decoding stage, even when reproducing a single virtual source, which results in wider sources [23] but reduced coloration effects, improved auralization of both diffuse and moving sources, and reverberant environments [19]. However, beyond the sweet spot area, significant spatial aliasing occurs, making HOA unsuitable when HAs are used in off-center locations [18,24]. Nevertheless, with proper trade-offs [24], HOA has proven to be a valuable choice for rendering VAEs, as confirmed by several perceptual studies [18,25]. It also offers advantages such as the easy collection of VAEs through spherical microphone array (SMA) recordings and their straightforward reproduction on an arbitrary speaker layout. Trade-offs to keep in mind include (i) the ambisonic order M—a higher M improves sound field reproduction accuracy, reducing localization errors and expanding the sweet spot [24], but an excessively high M can introduce coloration effects, which can still be mitigated by using a non-anechoic listening room [24]; (ii) the speaker array radius—larger radii can increase the sweet spot, but too large a radius requires a significantly higher M to maintain good localization accuracy [26]; (iii) the decoding technique—advanced decoders like AllRAD [27], with appropriate weighting choices, can enhance source width [28], localization accuracy [15,16,24], and spatial distribution [29].

Regarding the methods used to acoustically validate such complex reproduction systems—to understand how experimental results obtained in a VAE translate to the real world—multiple approaches have been followed, which can be broadly divided into physical and perceptual strands. Standard monaural and binaural room acoustical parameters [30] have been extensively employed to verify the system ability to generate a VAE that satisfactorily embeds the key physical acoustical properties of the target environment. Specifically, different studies have focused on reverberation time (T_{60}, T_{30}, T_{20}) [21,25,31–35], early decay time (EDT) [21,32,34,35], speech clarity (C_{50}, C_{80}) [21,25,32–35], definition (D_{50}) [35], center time (T_S) [35], sound strength (G) [21,35], interaural cross-correlation (IACC) [21,25,33,35], and the direct-to-reverberant ratio (DRR) [33,34]. Additionally, as predictor metrics of human localization performance, some studies have evaluated

binaural parameters such as the interaural level difference (ILD) and interaural time difference (ITD) [18,34,36]. Furthermore, a metric predicting human performance in speech intelligibility, the Speech Transmission Index (STI) [37], was also addressed in [21,33,35]. On the real perceptual assessment side, some works have performed and compared speech intelligibility tests with real subjects in both the VAE and the real environment, collecting speech reception threshold (SRT) measurements [25,32–34]. Similarly, other studies [31,38] have investigated and compared the perceived spatial location of sounds, including distance perception, between real and virtual environments (VEs).

Finally, it is also important to mention the role that the visual scene plays in VR for reproducing typical communication scenarios. The integration of a virtual visual environment (VVE) is crucial for enhancing the immersive experience, which, in turn, reinforces the auditory illusion [39]. Visual cues have been shown to significantly impact the realism of these environments [19], further encouraging near-real-life listener movements, which have been proven to influence key aspects such as sound source localization and speech intelligibility [40–43]. Speech intelligibility, in particular, is the crucial acoustical aspect that must be accurately assessed in educational environments, where effective communication is the primary goal. An all-encompassing audiovisual (AV) experience is vital for achieving the most accurate perceptual assessment of acoustics, closely mirroring the real room. For instance, contextual and source-related visual cues enhance the ability to accurately locate sound sources [36], while the visibility of the speaker's face and mouth movements greatly aids in speech comprehension [44,45].

In this context, VR visualization techniques are broadly divided into two types. The first, known as "window on world" (WOW) or Desktop VR, uses conventional monitors to display a 3D VE, allowing users to remain visually connected to their physical surroundings [46]. The second type, immersive VR, fully envelops the user's view within the virtual 3D environment, typically using a VR headset to create a more intense and immersive experience [47]. In WOW setups, increasing the number and size of displays can enhance immersion to a certain extent, but it may be more effective to use image projectors, as one projector can cover a larger area [19]. One WOW approach involves surrounding the user with projections on cylindrical canvases [48]. However, this setup requires significant calibration and expertise, and often leaves the floor and ceiling uncovered, resulting in an incomplete 360° experience [19]. A more advanced WOW system is the Cave Automatic Virtual Environment (CAVE), which projects images onto the walls, floor, and ceiling of a room-sized space to create a fully enclosed 3D environment [49,50]. Despite its immersive potential, the CAVE system is costly and presents acoustic challenges due to its flat, reflective surfaces [19]. On the other hand, VR headsets offer a different kind of immersion by using a wearable device that covers the user's eyes and ears, creating a fully immersive 3D digital environment [51]. This provides a highly personal experience and is now relatively affordable and easy to set up [19]. However, drawbacks include the bulkiness of the headset, potential acoustic distortions when using loudspeakers [19,52], and possible discomfort or disorientation from sensory isolation, such as dissociation from one's body or issues with spatial awareness [53,54]. Additionally, some users may experience cybersickness, a condition characterized by symptoms such as nausea, dizziness, and disorientation, which occurs due to a mismatch between visual and vestibular sensory inputs [55,56].

1.2. Aim of the Paper

In general, there is no VR technique that is universally superior; rather, the suitability of a method depends on the specific application in mind. Given this understanding, for the purpose of validating a VR methodology that can be widely applied for subjective evaluations of room acoustics—whether for existing spaces or those still in development—concerning the VAE reproduction, here, it was decided to focus on the flexible and well-established HOA technique. This choice was driven by the significant constraint on the number of available speakers and the need for easy placement, which is crucial for ensuring

the method's portability across various indoor environments. Third-order ambisonics (3OA), which requires 16 speakers, was selected as a trade-off. In particular, 3OA has been shown to result in lower perceived localization errors on the horizontal plane, that is, the most important for localization [12], compared to VBAP and MDAP when using an equal number of speakers [15]. Additionally, the choice of 3D 3OA strikes a balance between minimizing cost and complexity while still providing a satisfactory sense of immersion. This is also due to its strong capability in recreating diffuse and reverberant environments [19], which are precisely the types of educational spaces in which this VR methodology, for the environment design, is intended to be applied, with the aim of improving acoustical comfort and speech intelligibility. Regarding VVE reproduction, the VR headset-based method was chosen as a balanced solution, given its ease of setup and lower cost. Supporting this choice, other studies have shown that the impact on the acoustic field can be negligible, especially when weighed against the significant visual benefits it offers [36,57].

Thus, the purpose of this paper is twofold: (i) to present a low-complexity VR system based on well-established techniques that can be easily replicated in standard indoor spaces to reproduce virtual AV scenes and perform perceptual tests, such as speech intelligibility, learning effort, sound localization tests, and similar, (ii) to evaluate its adequacy for the perceptual assessment of acoustic quality and comfort in existing or work-in-progress environments, particularly by examining the virtual reproduction accuracy against a real educational space, such as a reverberant lecture room. The broader validation proposed combines significant metrics for the assessment of educational settings, incorporating both objective and subjective measurements. The objective measurements focus on standard monaural room acoustical parameters to perform a physical validation of the reproduced VAE. Particular attention is given to the speech-weighted C_{50} [58] and DRR parameters, which can be considered predictive metrics for human speech intelligibility performance within an environment [59,60]. Moreover, since binaural cues are crucial for both speech intelligibility and spatial sound localization [61,62], binaural parameters that strongly correlate with perceived human sound localization and spatial impression, such as ITD, ILD, and IACC, were also used as benchmarks, further involving their analysis during HA usage. To complement the objective investigation with a subjective assessment of the AV system, a group of students completed the Igroup Presence Questionnaire (IPQ) after experiencing the virtual AV reproduction of a scene emulating a typical university lecture room through the VR system. This approach allowed for the collection of ratings from individuals familiar with the real lecture room (which corresponded to the auralized version) regarding their sense of presence in the VE.

2. Materials and Methods

The validation procedure described in this paper pertains to the VR system introduced in the first part of this section and is organized into two main steps, referred to as intra-lab and inter-lab validation. The intra-lab validation focuses on (i) comparing real and virtual sound sources by analyzing differences in the binaural parameters both in the sweet spot and as one moves away from the center of the speaker array, and (ii) examining localization cue distortions when HAs are used as support devices during VAE reproduction. The inter-lab validation assesses the quality of the overall audiovisual reproduction using a real lecture room as a benchmark. First, the accuracy of the reproduced sound field is evaluated by comparing monaural and binaural room acoustical parameters measured in the real lecture room with those acquired in the laboratory during the auralization of the same lecture room. Next, the subjective assessment of the audiovisual experience is derived by reproducing the AV scene of a typical lesson in the lecture room through the proposed VR system and collecting subjects' responses to the IPQ.

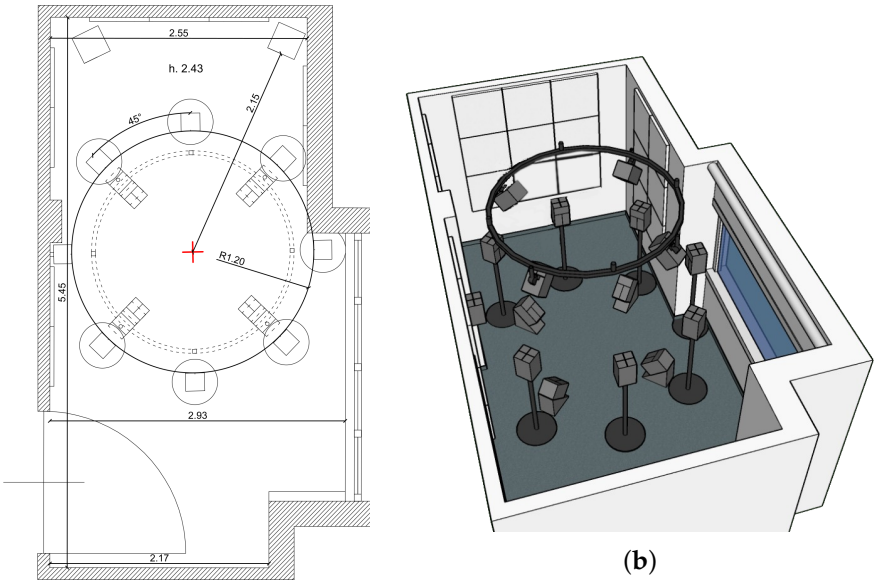
2.1. VR System

The VR system is installed in the Audio Space Lab (ASL) at the Politecnico di Torino. The ASL is a small listening room with a volume of 35.36 m^3, located on the first floor and overlooking an inner courtyard. The laboratory underwent acoustical treatment according to the room acoustical requirements outlined in the ITU-R BS.1116-3 standard for the subjective assessment of small impairments in audio systems [63]. The results of these treatments on the room's acoustical characteristics are detailed in [64]. Specifically, the room has reverberation times close to 0.17 s, which fall within the optimal range for octave band frequencies between 0.25 and 4 kHz. Additionally, background noise levels at the listening position range between noise rating (NR) 10 and 15 for frequencies up to 1 kHz and remain below 16 dB for the highest octave bands. The decision to set up the VR system in an acoustically treated room rather than an anechoic chamber is based on two main reasons. Firstly, to promote this VR system as a method for designing new spaces and assessing perceived acoustical quality and comfort, its implementation needs to be as simple and cost-effective as possible. Consequently, finding sufficiently large rooms and customizing them to build an anechoic chamber might hinder its widespread adoption. Secondly, some reflections can be beneficial in masking reproduction errors when precise objective assessments are not required [65]. The VR system is mainly made up of commercially available hardware and consists of two main components: (i) the VAE reproduction system, given by a 16.2 ambisonics setup; and (ii) the VVE reproduction system, given by a VR headset.

2.1.1. VAE Reproduction System

The VAE reproduction system consists of a spherical array of 16 Genelec 8030B two-way active monitors and two Genelec 8351A three-way active monitors, which are currently used for frequencies ranging from 30 to 90 Hz. These two monitors are placed on the floor, approximately 2 m in front of the listening area at the center of the speaker array. The 16 main speakers are equally distributed in space on a sphere with a radius of 120 cm, centered at a height of 121.5 cm from the floor. The radius was chosen to be as large as possible given the limited dimensions of the room. Specifically, the speakers are arranged in three rings at different elevation angles, i.e., −45°, 0°, and +45°, each having a different number of speakers. The upper and lower rings each contain four speakers, positioned to favor stereophonic listening. These speakers are spaced 90° apart and tilted so that their acoustic axes point toward the center of the sphere. The middle ring, which has the highest number of speakers to maximize spatial definition in the horizontal listening plane—where the listener experiences the greatest resolution in terms of spatial separation of sound sources [16]—hosts eight speakers that are spaced 45° apart. Figure 1 provides details on the speakers' placement, showing the ASL floor plan with the projections of the three speaker rings (Figure 1a) and the corresponding 3D model (Figure 1b).

The overall mounting system was conceived to (i) incorporate as much commercially available, easy-to-install mounting equipment as possible, and (ii) remain highly flexible to allow for easy placement and adjustment of the speakers. At the same time, it was designed to be as unobtrusive as possible to (i) minimize sound field distortions that could potentially lead to biased reproduction error metrics and perceptual test results [20], and (ii) limit the sense of constriction and occlusion that the listener might experience once inside the speaker array. Additionally, to ensure precise placement of the listener's head within the sweet spot, an adjustable chair was chosen to accommodate subjects of varying heights and to allow for rotations around the listener's longitudinal axis when the listening test permits.



(**a**) (**b**)

Figure 1. (**a**) ASL floor plan with the projected VAE reproduction system. The red cross marks the listening position. (**b**) A 3D model of the ASL with spherical 16-speaker array.

For driving the speakers, the 32-channel Antelope Orion32 sound card is used, which is powered by a high-end desktop PC equipped with an Intel® Core™ i7-12700F CPU and an NVIDIA GeForce RTX 3080 Ti GPU. The PC performs real-time signal processing at a sampling frequency of 48 kHz using a patch of the low-cost commercial Plogue Art et Technologie Bidule block-and-wire Digital Audio Workstation (DAW) to decode the 16-channel 3OA audio track into 18 distinct signals for the speakers. This processing particularly involves blocks from the open-source IEM plug-in suite [66] developed by the Institute of Electronic Music and Acoustics (Austria). Specifically, for decoding the ambisonic signal, the AllRADecoder block is used, configured to decode a 3OA signal with SN3D normalization using the All-Round Ambisonic Decoding (AllRAD) technique [27]. Although AllRAD was initially designed for ambisonic decoding on irregular speaker setups, it was still selected for this regular system since it is a very efficient strategy, holding for regular setups as well. Moreover, since the HOA rendering relying on the basic weighting technique can physically reconstruct the correct sound field in anechoic conditions up to a maximum cut-off frequency limit [16]:

$$f_{\mathrm{HOA}} = \frac{343\ \mathrm{m/s}}{2\pi r} M, \tag{1}$$

which is equal to 1.9 kHz for a sweet spot radius (r) of 8.5 cm (approximately equal to the average radius of the human head) and decoding order (M) equal to 3, the max-rE weighting is used to adjust the decoded signals to perceptually focus the high-frequency energy toward the expected direction, further improving the reconstruction of ILD [67]. However, this comes with the trade-off of slightly undermining the reconstruction accuracy at lower frequencies, but it still remains perceptually valid. The choice of max-rE weighting also helps with coloration error issues, which are minimized independently on the ambisonic order and possible off-centerings of the listener's position [24]. Finally, max-rE represents the weighting leading to lower localization errors also in the case of listeners placed out of the sweet spot region [15,24]. Since the loudspeakers are more than 1 m away from the listening position, no near-field compensation is applied.

Furthermore, in order to maximize the sound field reproduction accuracy, a tuning procedure was performed, by synchronizing the acquisition of a class-1 omnidirectional microphone (flat frequency response from 5 to 20,000 Hz) placed in the center of the speaker

array. In this way, filters and time delays to be applied to the speaker signals were derived to flatten the frequency response within the sweet spot from 90 to 10,000 Hz (and from 30 to 90 Hz through the subwoofers) and temporally align the arriving signals. The analysis focused on the 90 to 10,000 Hz frequency range, as it encompasses the speech frequency content [68], which is the primary frequency range involved in typical auditory scenarios within educational environments.

Particularly, all speakers on the sphere were high-passed in frequency using a fourth-order filter with a cut-off frequency at 60 Hz and 12 dB/oct smoothing to create a Linkwitz–Riley crossover through the multichannel equalizer IEM MultiEQ plug-in. Moreover, for the individual speaker tuning, band-pass IIR filters were used with quality factor $Q < 5$ to flatten the frequency response of all speakers, bringing them as close as possible to each other (smoothed in 1/3 octaves). The two speakers for the lower frequencies were also equalized to ensure proper balance with the 16-speaker array. Next, using common IIR filters on all channels, the system was equalized so as to correct for spectral cancellations and emphases due to the simultaneous use of multiple speakers. The final frequency response of the system given by the fine-band frequency spectrum of each of the 16 speakers in the center of the speaker array is shown in Figure 2a.

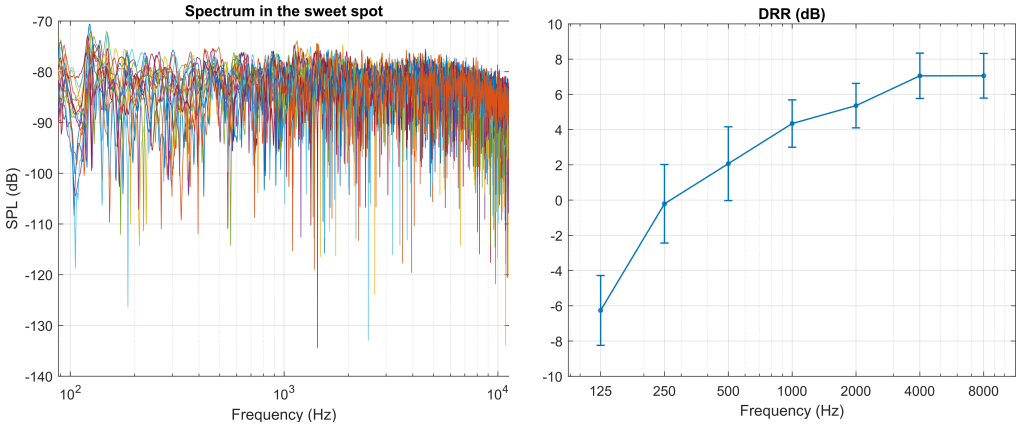


Figure 2. (**a**) Normalized spectrum of the 16-speaker array frequency response in the listening center from 10 to 10,000 Hz. (**b**) Average and standard deviations (SDs) of DRR values computed in the center of the 16-speaker array from 125 to 8000 Hz octave bands.

Although the fine-band spectrum for each speaker is characterized by several very narrow peaks, they generally compensate each other, except around 105 Hz, where almost all speakers show a dip due to the presence of a room mode. This could lead to discrepancies between the target sound field and the virtually reconstructed one at those frequencies. This observation is further supported by the analysis of the DRR calculated for the 16 speaker signals arriving at the sweet spot. In the case of a correctly reconstructed VAE, the DRR should exhibit positive values, indicating that the laboratory room reflections do not interfere with the generated sound field. Figure 2b shows the DRR analysis (average and SD values) in octave bands from 125 to 8000 Hz, performed using a MATLAB script based on a function from the open-source MATLAB toolbox in [69]. This function determines the direct sound as the peak of the squared impulse response and returns the DRR value using a 5 ms time window centered on the peak to select the direct sound [70]. As expected from the spectrum analysis, the DRR is characterized by negative values, around -6 dB, for the 125 Hz octave band, which corresponds to the frequency at which the room mode occurs.

Finally, concerning the temporal alignment of the signals at the center of the speaker array, all delays are compensated, even though this might result in small head movements outside the sweet spot region causing strong phase variations, which can lead to coloration and localization errors [71]. This decision was made after subjective listening trials, in which

introducing differences in signal arrival times did not improve either the localization error or the coloration effects when the listener was slightly off-center. Moreover, coloration effects are also mitigated by the usage of a non-anechoic listening room [72].

2.1.2. VVE Reproduction System

The VVE reproduction system consists of the Meta Quest 2 VR headset, directly connected via Meta Quest Link to the same high-end PC that drives the VSE reproduction. The visual scene is presented as a simple stream of either a pre-recorded video within a real environment or a video of a synthetic rendering of an environment. The video streaming is handled through a custom application created using Unreal Engine by Epic Games [73], running on the PC. This application manages the selection of the video to be reproduced and performs the video streaming on the headset. Specifically, the application handles the playback of 360° stereoscopic videos with a resolution of 3840 × 3840, a frame rate of 30 fps using the H.264 codec.

2.1.3. Software Framework for the Overall AV Playback

The Unreal application for visual scene streaming and the Bidule patch for audio signal processing work in parallel as back-end components of the software framework devised for AV playback. A front-end application, developed using MATLAB, is responsible for (i) handling the selection of the AV scene—potentially including an associated listening test; (ii) triggering the synchronized AV reproduction; and (iii) collecting possible outcomes from the selected AV scene, such as speech intelligibility test results or subjects' comments on perceived acoustic comfort. To successfully run the AV reproduction, all three applications exchange real-time messages through the Open Sound Control (OSC) protocol. Figure 3 illustrates a schematic of the software framework, showing the OSC exchanges between the different applications.

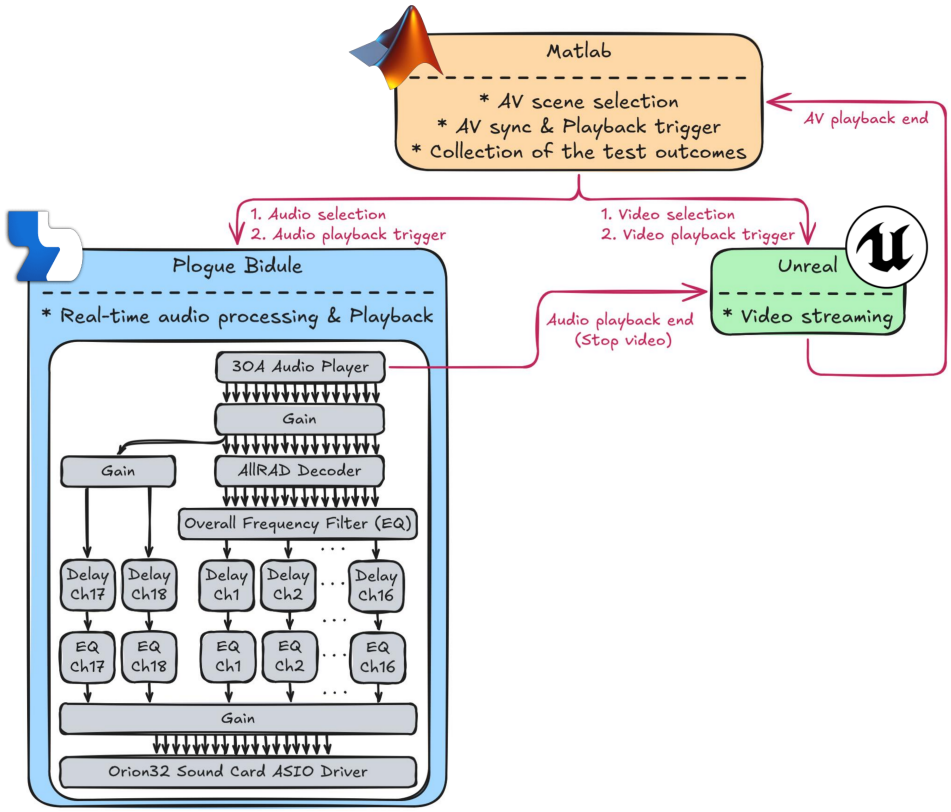


Figure 3. Scheme of the software framework components and their interconnections as the basis of the proposed VR system.

Overall, the budget required for the installation of the entire system (including construction, electronic hardware, and software components), excluding the room's acoustic treatment—which strongly depends on the structural characteristics of the available space—amounts to approximately EUR 20,000, which aligns with the average budget spent by medium-sized architectural firms on multi-year subscriptions for commercial software used for 2D and 3D technical drawing, graphic design, and room acoustic simulation tools. Figure 4 shows a picture of the ASL with the VR system during a speech intelligibility test on a subject. A more detailed explanation of the specific hardware, mounting construction, loudspeaker placement, and software management for the audio system tuning procedure can be found in [74].



Figure 4. Picture of the VR system inside the ASL during the execution of a speech intelligibility test on a listener.

2.2. Intra-Lab Validation

The goal of the intra-lab validation was to analyze the behavior of the system within the sweet spot and at progressively off-center positions. To achieve this, binaural room impulse responses (BRIRs), acquired using both real sound sources (RSSs) and virtual sound sources (VSSs), were compared based on the computed ILD and ITD values. This analysis was performed twice to assess how the results changed when HA microphones were active during listening.

Measurement Procedure

For both RSSs and VSSs, 16 sound source positions were selected corresponding to the locations of the speakers in the spherical array. These positions include eight equally spaced locations starting from the 0° azimuth angle on the middle ring at 0° elevation, and four equally spaced locations starting from the +45° azimuth angle on both the upper and lower rings at +45° and −45° elevation angles, respectively. RSSs were created by directly driving one speaker from the spherical array at a time, while VSSs were generated using 3OA virtual panning at the selected locations through the IEM MultiEncoder plug-in, instantiated within the Bidule patch used for handling VSE reproduction. Two-second exponential sweep signals ranging from 90 to 20,000 Hz at a 48 kHz sampling frequency were used as excitation signals. The BSU Head and Torso Simulator (HATS) from HEAD acoustics (with a 48 kHz sampling frequency and equal frequency response between the left and right ears from 22 to 9000 Hz) was used to acquire the BRIR twice—once with and once without behind-the-ear HAs (portable hearing lab headsets from BatAndCat Sound Labs [75]), as shown in Figure 5. For recording the HATS signals, the proprietary HEAD acoustics SQobold mobile front-end was directly connected to the HATS as a data logger. For recording the HA signals, a notebook PC running a customized Bidule patch was connected to the Roland UA-1010 Octa Capture sound card, sampling at 48 kHz.

Three HATS positions, each with a different off-center distance from the center of the speaker array, were selected to enable direct comparison with the measurements in [18]: 0, 10, and 20 cm. In total, $16 \times 2 \times 2 \times 4 = 256$ BRIRs were computed, on which ILD and ITD values were evaluated. MATLAB scripts were used to calculate both ILDs and ITDs based on the *ita_roomacoustics_IACC* function from the open-source ITA Toolbox MATLAB library [76].

The ILD was computed as the energy ratio between the left and right ear signals, band-passed from 1 to 9 kHz. The low cut-off frequency was selected based on [77], as this is the frequency at which ILD cues begin to impact sound localization perception due to the head shadow effect. Following the approach in [18], the ILD analysis results are presented as average errors between the RSSs and VSSs across all source positions:

$$\Delta\text{ILD} = \sum_{i}^{N} \frac{|ILD_{RSS_i} - ILD_{VSS_i}|}{N}, \tag{2}$$

where N equals 16, representing the total number of sound source locations used. To provide a clearer picture of the contribution of elevated sound sources to the overall error, two additional ΔILD values are presented: one including only the eight sound sources placed at a 0° elevation angle, and another including only the eight sound sources at +45° and −45° elevation angles.



Figure 5. BRIR measurements within the ASL, with the HATS wearing the HAs placed at the 0 cm off-center position from the sweet spot.

The ITD was obtained as the lag of the peak of the IACC function of the left and right ear signals, band-passed from 90 to 1400 Hz. The high cut-off frequency was selected based on [77], beyond which humans become insensitive to ITD cues. Although the ITD parameter is sensitive to frequency variation [78], a single-valued ITD is considered, as the variability in ITD with frequency has been shown to be of little perceptual importance for sound localization [61]. Since (i) ITD cues are not used to discern the elevation of sound sources [77], and (ii) the degree of human sensitivity to ITD cues varies according to the specific azimuth angle—being highest for sounds from the frontal direction and lower as sound lateralization increases—the ITD results are presented individually for both the VSSs and RSSs for each sound source located on the middle ring.

2.3. Inter-Lab Validation Against a Real Lecture Room

The goal of the inter-lab validation is to analyze the reliability of the system in recreating existing environments, thus assessing how well the results from objective and subjective

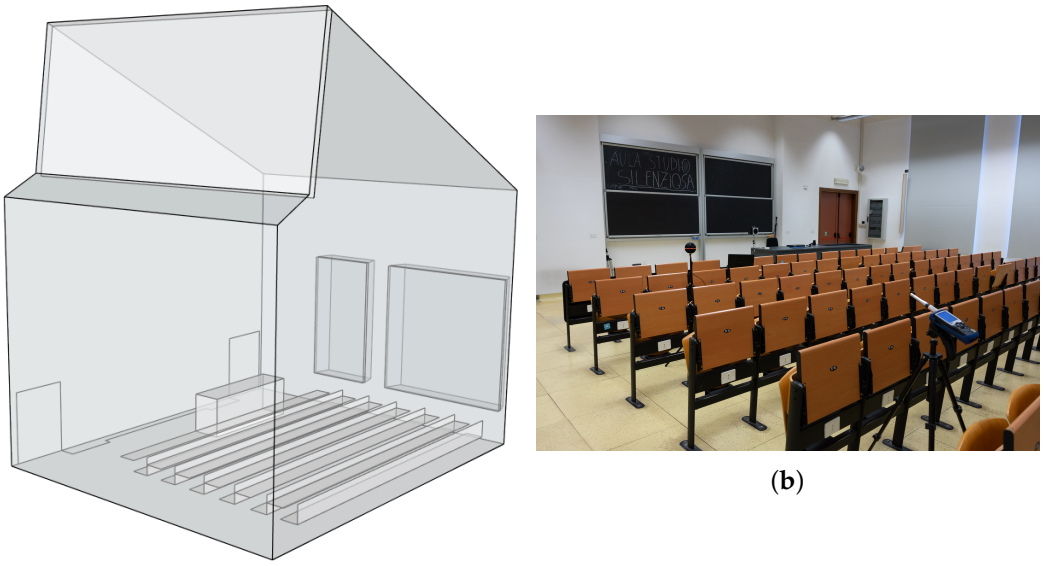
metrics evaluated within the VAE translate to the real world, using a real educational environment as a benchmark. To achieve this, the inter-lab validation was divided into two phases: the objective acoustical evaluation and the subjective audiovisual evaluation.

During the former, spatial RIRs were acquired inside a real lecture room, reproduced through the proposed VAE reproduction system, and recorded again within the ASL. This process resulted in both monoaural and binaural RIRs obtained from both the real and virtual lecture rooms, allowing for the computation and comparison of monoaural and binaural acoustic parameters. Additionally, the same procedure was performed twice to check what happens to the binaural parameters in the case of HA microphone usage.

During the latter phase, focusing on a purely subjective assessment standpoint, 360° AV recordings simulating a lecture were filmed inside the real lecture room and played back within the ASL. Responses to the IPQ questionnaire on the sense of presence, elicited during the virtual experience, were collected from a pool of students who were accustomed to attending lectures in the real lecture room.

2.3.1. Case Study

The lecture room selected as the case study is an 800 m^3 room at the Politecnico di Torino. It is located on the ground floor of a renovated building surrounded by a private pedestrian area, primarily used by university students during class intervals. The lecture room is characterized by an irregular ceiling with an average height of 7.8 m, an acoustically reflective linoleum floor, two windows, and two doors. Inside the room, there are six rows of 15 wooden seats with tables, arranged in front of a blackboard and a large acoustically reflective desk, behind which the lecturer typically stands. Additionally, the room lacks acoustic treatments, as all walls and the ceiling are finished in plaster. The room has an average T_{30} of 2.3 s from 250 to 4000 Hz, an average Speech Transmission Index for Public Addresses (STIPA) value of 0.57, and an A-weighted global equivalent background noise level of 43.3 dB. These factors create very challenging conditions for typical real-life communication situations where speech understanding is crucial. According to [79], the optimal values for T_{30}, STI, and background noise level should be 0.8 s, greater than 0.6, and less than 41 dBA, respectively. Figure 6 shows the 3D model and a photograph of the environment, while Figure 7 illustrates the room floor plan.



(**a**) (**b**)

Figure 6. (**a**) A 3D model of the lecture room. (**b**) Picture of the lecture room taken with the same orientation as the 3D model on the left.

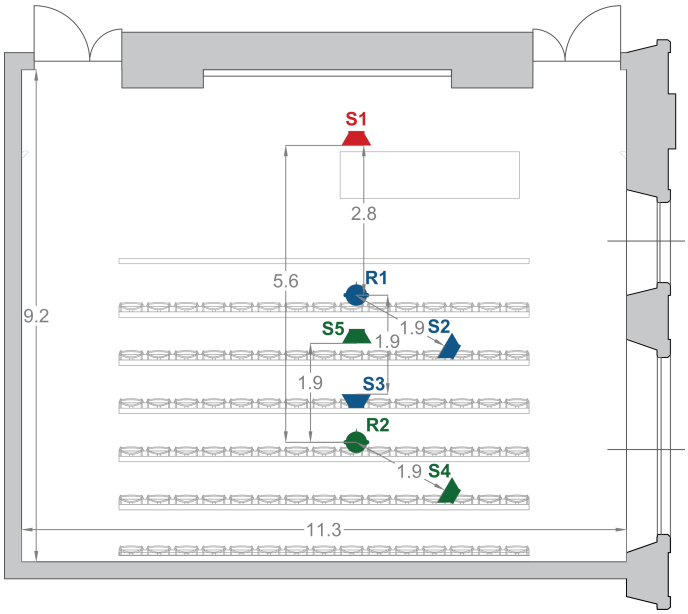
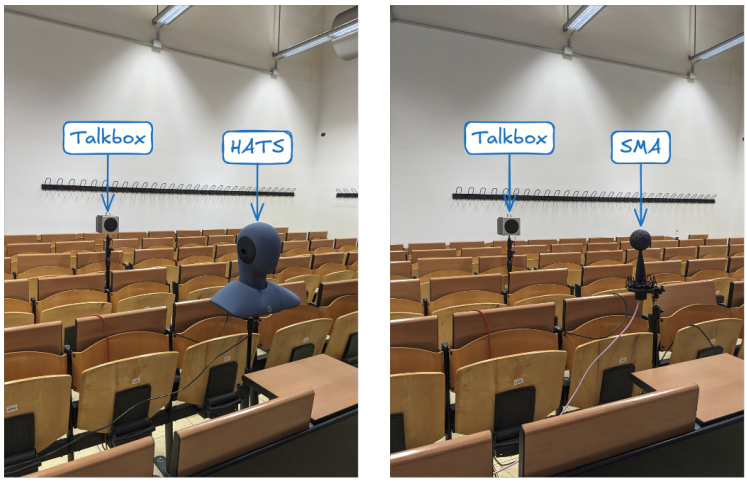


Figure 7. Floor plan of the lecture room with sound sources (S1, S2, S3, S4, and S5) and receiver positions (R1 and R2), composing the auditory scenes used for the measurements of ILD, ITD, IACC, DRR, speech-weighted C_{50}, and DRR. Sources and receivers of the same color identify all source–receiver pairs used for the RIR acquisition, except for S1, which was used as a sound source for both R1 and R2, resulting in a total of 6 RIRs.

2.3.2. Objective Acoustical Evaluation

In order to evaluate standard monoaural room acoustics parameters such as T_{20}, T_{30}, EDT, C_{50}, C_{80}, D_{50}, and T_S, according to the standards BS EN ISO 3382-1 [30] and 3382-2 [79] for measurements of room acoustic parameters inside performance spaces and for reverberation time in ordinary rooms, nine RIRs were acquired inside the lecture room three times each, with each corresponding to a different spatial configuration for the source–receiver pair. In particular, two sound source positions were defined, both at a height of 1.5 m from the floor, representing typical locations occupied by the main speaker inside the room. Each position was associated with a different set of receiver locations, uniformly distributed to cover the entire audience area. The room impulse response (RIR) measurements were performed with the heating, ventilation and air conditioning (HVAC) system off, doors and windows closed, curtains drawn, and the video projector active, in unoccupied conditions, using the backward-integrated impulse response method. The Brüel and Kjær 4292-L omnidirectional sound source, compliant with the requirements of [30], was used as the source, driven by the Lab Gruppen LAB300 amplifier, emitting a 5-second exponential sweep at a 48 kHz sampling frequency. The NTi Audio XL2 omnidirectional class-1 sound level meter (SLM) was used as the receiver at a height of 1.2 m from the floor, along with the Eigenmike em-64 SMA for spatial RIR acquisition to be reproduced inside the ASL. The 64-channel signals captured through the SMA at a 48 kHz sampling rate, connected to a customized Bidule patch, were converted in real-time into fifth-order ambisonics signals (5OA) using appropriate filters obtained through an SMA spatial calibration performed according to the procedure in [80]. The 5OA signals obtained in this way were played back through the VSE system and recorded by placing the SLM at the center of the speaker array. All acoustical parameters were then evaluated using a MATLAB script based on the *ita_roomacoustics* [81] function, that is a collection of validated routines from the ITA Toolbox library used to compute all standard room acoustics parameters starting from impulse response measurements following the methods described in the standards BS EN ISO 3382-1. Both the real and the virtual lecture room values and are provided as spatial averages in octave frequency bands from 125 to 8000 Hz.

Concerning the RIRs used to compute monaural and binaural parameters that significantly influence speech intelligibility—specifically, speech-weighted C_{50}, speech-weighted DRR, ILD, ITD, and IACC—they were acquired using source–receiver spatial configurations that match typical communication scenarios inside the lecture room, as these parameters are location-dependent. A total of six source–receiver pairs were selected, with the source and receiver fixed at 1.5 m and 1.2 m, respectively, from the floor, as illustrated in the floor plan shown in Figure 7. To emulate the directivity of the human voice, the NTi Talkbox acoustic signal generator (which has a flat frequency response from 100 to 10,000 Hz) was used as the sound source, emitting a 5 s exponential sweep signal at a sample rate of 44.1 kHz. At each receiving position, a set of different receivers was used, the SLM, the SMA, and the HATS, with and without the HAs, as explained in Section 2.2. Figure 7 shows the measurements inside the real lecture room for both the HATS (Figure 8a) and the SMA (Figure 8b). The obtained 5OA signals were again reproduced within the ASL, acquiring six additional RIRs using both the SLM and the HATS, with and without the HAs. To further verify the sweet spot extension when a real environment is auralized, BRIR recordings were repeated three times, each with a different off-center position of the recording equipment relative to the center of the speaker array, specifically at distances of 0, 10, and 20 cm. Both speech-weighted C_{50} and DRR, obtained as described in Section 1.1, were computed for each source–receiver spatial configuration as a frequency average from 250 to 4000 Hz octave bands, following the method in [58]. The DRR is also provided as a broadband value, ranging from 125 to 8000 Hz octave bands. ILD, ITD, and IACC were computed as stated in Section 2.2, after properly truncating the BRIRs to exclude background noise in the tail. For ITD, the early part of the BRIR was considered, as, for a fixed source–receiver spatial configuration, the perceived localization of a sound source in a real room mainly depends on the direct sound and the first reflections [82]. Regarding the IACC, which provides cues on spatial aspects beyond just localization perception, it was divided into early and late components to account for the perceived source width and the perceived sound spatial envelopment [83], respectively. Both early and late IACC were obtained from the BRIR signals as a spectral average across 500 to 2000 Hz octave bands [84].



(**a**) (**b**)

Figure 8. Acoustical measurements inside the real lecture room, in case of (**a**) 5OA RIR and (**b**) BRIR.

2.3.3. Subjective Audiovisual Evaluation

In order to perform the perceptual assessment of the AV reproduction quality of the VR system, a scene emulating a lecture in the real lecture room was recorded. The scene featured a main speaker who spoke and walked around the area typically occupied by the lecturer, as well as a group of students distributed throughout the audience who listened, engaged in small talk, and even asked questions to the lecturer. A 360° AV recording

setup [85], consisting of an Insta360 Pro 360° camera mounted on top of a Zylia ZM-1 19-capsule SMA (see Figure 9a), was placed at the listener's location, which was in the middle of the second row at a height of 1.2 m from the floor. The video was captured using the camera, controlled via the Insta360 proprietary Android application on a smartphone, while the audio was recorded at a 48 kHz sample rate by connecting the SMA to a notebook PC running a customized Bidule patch. This patch converted the 19-channel signal into a 3OA signal using the Zylia Ambisonics Converter plug-in. After post-production, the video was exported in H.264 format, resulting in a three-minute .mp4 file containing a 4K 3D 360° video with a resolution of 3840 × 3840 pixels and a frame rate of 30 fps. Figure 9b shows the equirectangular preview of the recorded scene.

To run the virtual experience through the VR system and perform the perceptual AV assessment, both the video and a three-minute excerpt of the 3OA audio track, corresponding to the exported video, were loaded into the VR system software. A total of 26 self-reported normal-hearing (NH) subjects (18 male, 7 female, and one who preferred not to declare a gender) aged 22 to 39 (average: 26.5, SD: 4.8) were recruited voluntarily from the students who regularly attended lessons in the lecture room under investigation. After the test, subjects were rewarded with pens, candies, water bottles, and notepads. None of them reported vision issues except for prescribed corrective glasses, which could be worn during the experiment using the adapter available for the headset. Additionally, none of the subjects declared any history of epilepsy or other clinical conditions that might have interfered with the experience or caused physical or psychological harm due to the immersive reproduction. Approximately 62% of the subjects reported having little to no experience with immersive AV reproduction systems, while 38% claimed to have more than medium- to expert-level experience. Four subjects reported experiencing cybersickness during the AV reproduction. Soon after experiencing the immersive AV scene, all subjects were administered with an Italian adaptation of the IPQ via a tablet.



(**a**) (**b**)

Figure 9. (**a**) Setup of the AV recording equipment. (**b**) Equirectangular preview of the lecture scene inside the lecture room.

The IPQ provides a scale for measuring the sense of presence experienced in a VE [86]. All questions are divided into three subscales, plus an additional general item regarding the overall sense of presence, referred to as the 'sense of being there'. The subscales are (i) spatial presence, which relates to the sense of being physically present in the VE; (ii) involvement, which measures the attention paid to the VE and the level of engagement experienced; and (iii) experienced realism, which refers to the subjective experience of realism in the VE. For the purpose of this study, a total of 11 questions were selected from the IPQ, excluding those that could not be applied to the investigated VR system, such as questions about the sense of acting in the VE that are relevant only for interactive systems, and questions that were found to be difficult to understand in pilot tests. The table with both the English and Italian versions of the selected questions is provided in Appendix A.

Moreover, to gain a clearer understanding of the separate contributions of the audio and video renderings to the overall realism, two additional questions were added, each focusing on a different aspect of the VE. The question on experienced realism (item 13) was re-proposed to address, in one case, the realism of the VAE (item 13a*) and, in the other, the realism of the VVE (item 13b*). To ensure unbiased translations, the Italian version of the questions was generated using the ChatGPT artificial intelligence tool. Although the IPQ is typically used to compare different conditions, the goal of this work was to assess whether subjects could envision themselves in the real situation based on the VEs. This was assumed to be true if subjects responded positively on average to the items related to presence, involvement, and realism. The IPQ response scale ranges from −3 to +3, where −3 is 'very bad', indicating a very poor experience that differs from real life, 0 is neutral, and +3 is 'very good', representing an excellent experience that resembles real life. Average values were computed, grouping all questions belonging to the same subscale. Additionally, the analysis on the average for the added questions on the experienced realism (the auditory, the visual and the auditory–visual one) is reported separately.

Finally, an open interview was conducted to gather opinions and comments on the experience, facilitating the identification of common themes that could highlight critical issues or strengths of the AV reproduction not addressed in the questionnaire.

3. Results

The results presented below for the objective metrics are analyzed both in absolute terms and in relation to the threshold of perception for each parameter (just-noticeable difference, JND) [35]. This approach aims to correlate the absolute differences between real and virtual sounds with the audible discrimination threshold, which is commonly used as a reference in room acoustics evaluation [35,36].

3.1. Intra-Lab Validation

Figure 10a shows the average and SD values of the absolute ILD difference, ΔILD, between Real RSS and VSS across all 16 sound source positions for both the HATS and HA measurements as a function of the off-center distance from the center of the speaker array (i.e., 0, 10, and 20 cm). Additionally, the JND value from [36] is included for reference. Figure 10b,c present the same analysis, but restricted to the eight sound sources on the middle ring and the eight sound sources with elevation angles of either +45° or −45°. In the case of the HATS at 0 cm off-center, both the ΔILD average and the SD across all 16 sound sources fall within the JND. For the other off-center positions, the average values worsen as expected, reaching a maximum of 2 dB at a 10 cm distance, while the SDs exceed the JND. Notably, at 20 cm off-center, the sound sources that primarily contribute to the increased errors are those on the middle ring, while for the elevated sound sources, both the average and SD remain below 1 JND. When using the HA microphone, the average at 0 cm off-center is slightly higher than that of the HATS but still below the JND, while the SD exceeds the JND by around 1 dB. The results for the other off-center positions follow the same trend as observed with the HATS.

Figure 11a,b compare the ITD values between RSSs and VSSs for all off-center positions in the HATS and HA cases, respectively, along with the corresponding JND. The values are presented as a function of sound source location, as the JND of the ITD increases with sound source lateralization [36,87,88], ranging from about 10–20 μs to 120 μs for broadband signals. This implies that there is no single value that can be used for comparison with an averaged ITD error. The analysis is performed exclusively for the eight sound sources on the middle ring, as the azimuthal plane is the only plane in which ITD cues are effective for discriminating sound directions [77]. Looking at the trends obtained for both the HATS and the HA, the VSS closely follows the same pattern as the RSS. More specifically, in the HATS case, VSSs at 0 cm off-center remain within the JND for all sound source positions. However, similar to the ILD, as the off-center distance increases, the ITD exceeds the JND for certain specific positions. As expected, this discrepancy occurs for sound directions on

or near the median plane. A similar trend is observed in the case of HA usage, where some very slight deviations outside the JND are found even for VSSs at 0 cm off-center.

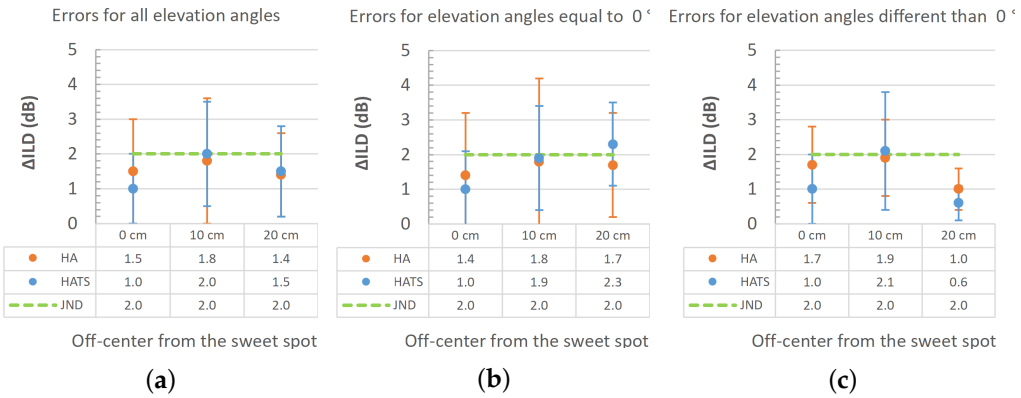


Figure 10. Average and SD values, along with JND values of ΔILD, band-passed from 1 to 9 kHz, between RSS and VSS for both HATS and HA measurements, as a function of off-center distance, across (**a**) all 16 sound source positions; (**b**) all 8 sound source positions on the middle ring; (**c**) all 8 sound source positions at either +45° or −45° elevation angles.

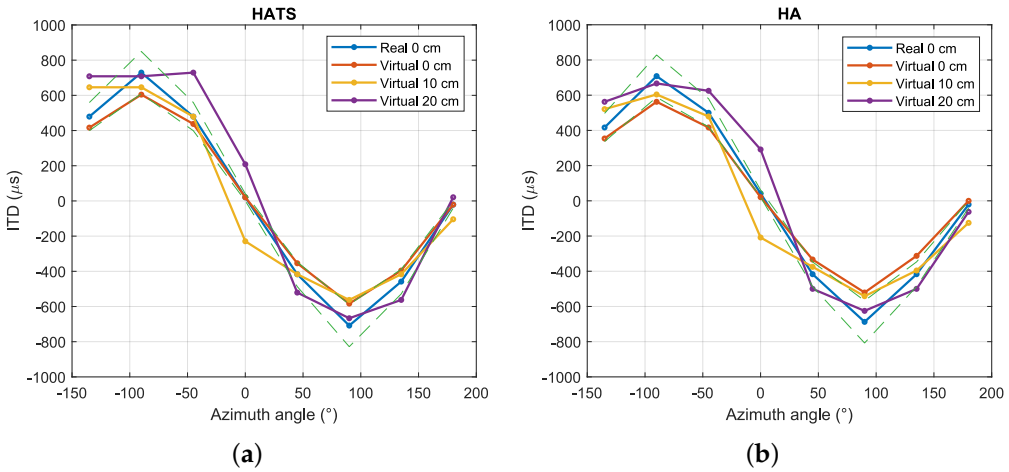


Figure 11. ITD values, band-passed from 90 to 1400 Hz, for RSSs and VSSs on the middle ring and JND limits for all off-center positions as a function of sound source location for measurements with (**a**) the HATS; (**b**) the HAs.

3.2. Inter-Lab Validation Against a Real Lecture Room

3.2.1. Objective Acoustical Evaluation

Figure 12 shows the comparison between the real and virtual lecture rooms based on standard monaural room acoustical parameters, i.e., EDT, T_{20}, T_{30}, D_{50}, C_{50}, C_{80}, and T_S.

Specifically, spatial averages and SD values are plotted as a function of octave-band frequencies, along with the corresponding JND values [30]. The absolute values and trends of the parameters for the real lecture room align with expectations for a highly reverberant space, showing high values for the temporal parameters and low values for the energetic parameters at mid-frequencies, where the contribution to speech intelligibility is most significant. For all temporal parameters, the averages and most of the SDs are within the JND, except for EDT in the 125 Hz octave band, where the virtual lecture room presents an average value lower than the real one, exceeding the JND by 0.05 s. This suggests that the virtual room has a little deficit in recreating the very first reflections. This observation is further supported by the average values of D_{50} and C_{50} at the 125 Hz octave band, where the difference is more pronounced, i.e., 9.4% and 2.2 dB, respectively. However, for octave bands ranging from 250 to 4000 Hz, which are more significant for speech comprehension,

the virtual lecture room shows average energetic parameters that fall perfectly within the JND, with comparable SDs to those of the real lecture room.

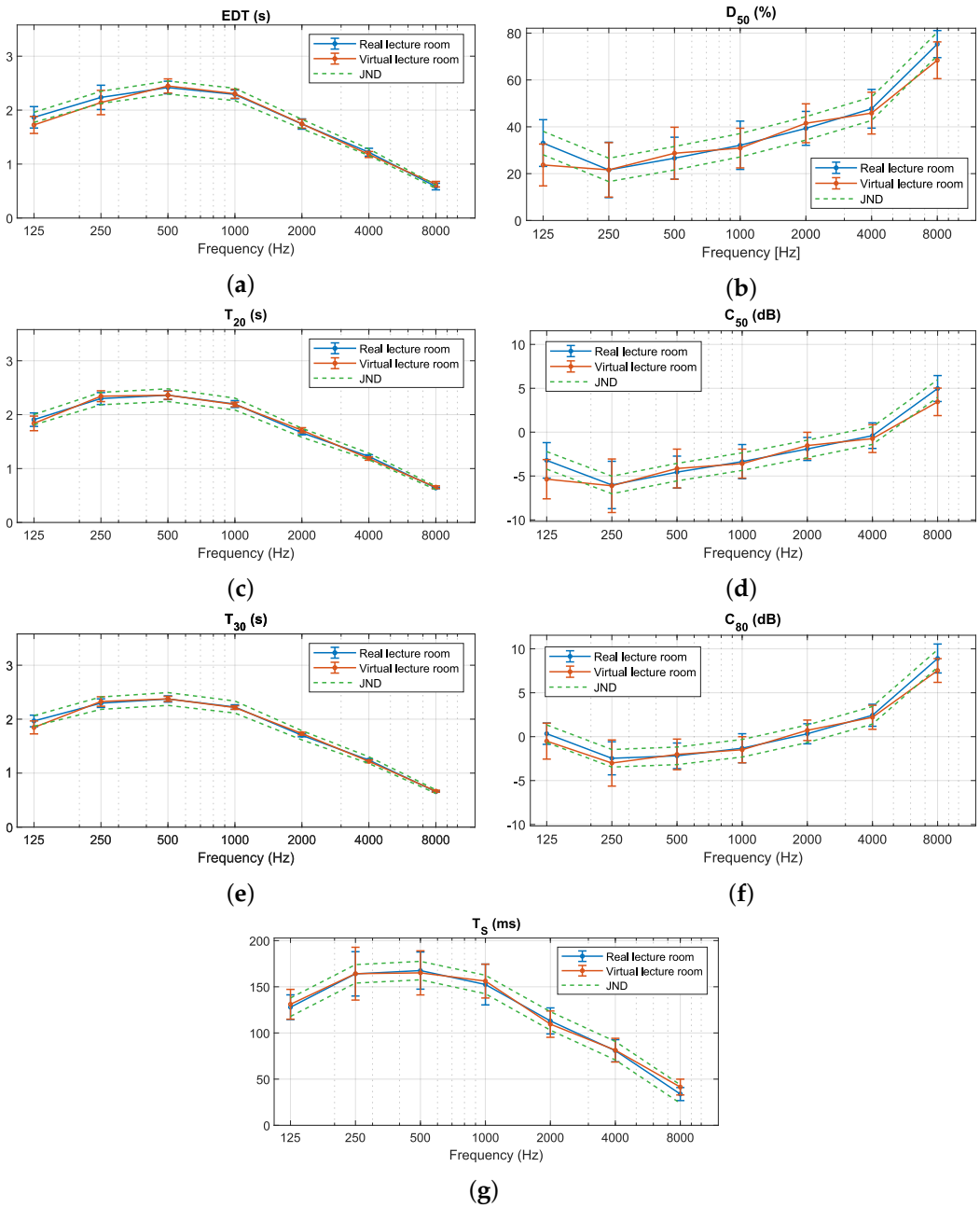


Figure 12. Spatial averages, SD, and JND values in octave bands from 125 to 8000 Hz for the real and virtual lecture rooms for (**a**) EDT; (**b**) D_{50}; (**c**) T_{20}; (**d**) C_{50}; (**e**) T_{30}; (**f**) C_{80}; (**g**) T_S.

Figure 13 presents the analysis of source–receiver configuration-dependent parameters, such as DRR, which are also related to speech intelligibility prediction, including speech-weighted C_{50} and DRR, along with the corresponding JNDs taken from [30,70].

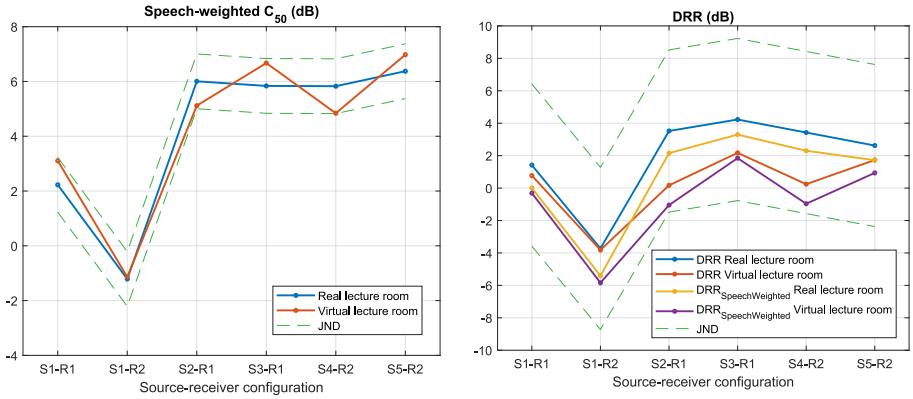


Figure 13. (**a**) Speech-weighted C_{50} values averaged across the 500 to 4000 Hz octave bands and JND limits for the real and virtual lecture rooms as a function of the source–receiver configuration. (**b**) DRR values band-passed from the 250 to 8000 Hz octave bands and speech-weighted DRR values averaged across the 500 to 4000 Hz octave bands for the real and virtual lecture rooms as a function of the source–receiver configuration.

As expected, in the real lecture room, the highest values for all parameters are observed for the source–receiver pair with the shortest distance, while the worst value is found for the configuration with the greatest source-to-receiver distance (S1-R2). All parameters computed for the virtual lecture room fall within the JND compared to the real lecture room for all source–receiver locations.

Figure 14 shows the comparison between the real and virtual lecture rooms based on ILD and early ITD values, for both HATS and HA measurements. The analysis is presented as a function of the source–receiver configurations, along with the corresponding JNDs. For ITD, since the BRIRs were acquired in a highly reverberant environment, the reported JND refers to the value found in [89], where an ITD JND of about 140 µs was measured from both guitar and violin music signals inside a lecture hall with a reverberation time of 1.7 s. For both the real and virtual lecture rooms, the ILDs exhibit the expected negatively increasing pattern with angular separation in the horizontal plane, while the radially spanning positions (S1-R1 and S1-R2) show only subtle changes in ILD, possibly due to early reflections. However, while HA measurements inside the virtual lecture room fall within the JND compared to the real room for all source–receiver configurations, for the HATS measurements, the S3-R1 pair in the virtual room deviates from the real room, exceeding the JND of less than 1 dB. Similarly to the ILD, the early ITD shows an increasing pattern with angular separation in the horizontal plane and only subtle changes for positions spanning radially. In the case of HATS measurements, all ITD values for the virtual lecture room lie within the JND limits, while in the case of HA measurements, the ITD for the virtual lecture room very slightly exceeds the JND in configurations with greater angular separation. This is likely due to the previously noted difference in early reflections at very low frequencies between the real and virtual rooms.

Figures 15 and 16 show the comparison between the virtual and real lecture rooms based on early and late IACC values, respectively, for both HATS and HAs, as the source–receiver configuration varies, along with the corresponding JNDs [30].

In the real lecture room, higher early IACC values are found for sound sources located on the median plane of the receiver, except for the S1-R2 pair, where the decorrelation between the left and right ears increases due to the lower DRR value, as the source is farther from the receiver compared to all other configurations. The late IACC, being only influenced by late reflections, is location-independent, and thus, shows relatively consistent values across all source–receiver configurations. Moreover, the values are quite low, which, consistent with the high T_{30} observed, would result in a strong sense of sound envelopment for the listener. In the virtual lecture room, all early IACC values fall within the JND

compared to the real room, except for configurations S1-R1 and S5-R2 in the HATS case, and configurations S1-R1, S2-R1, and S5-R2 in the HA case. In all cases where the IACC does not comply, the measured value is lower than the desired one. For the late IACC, a mirrored trend is observed, where the values are generally perceptually higher than desired for most source–receiver configurations. The only exceptions are for configurations S1-R2 and S5-R2 in both the HATS and HA cases.

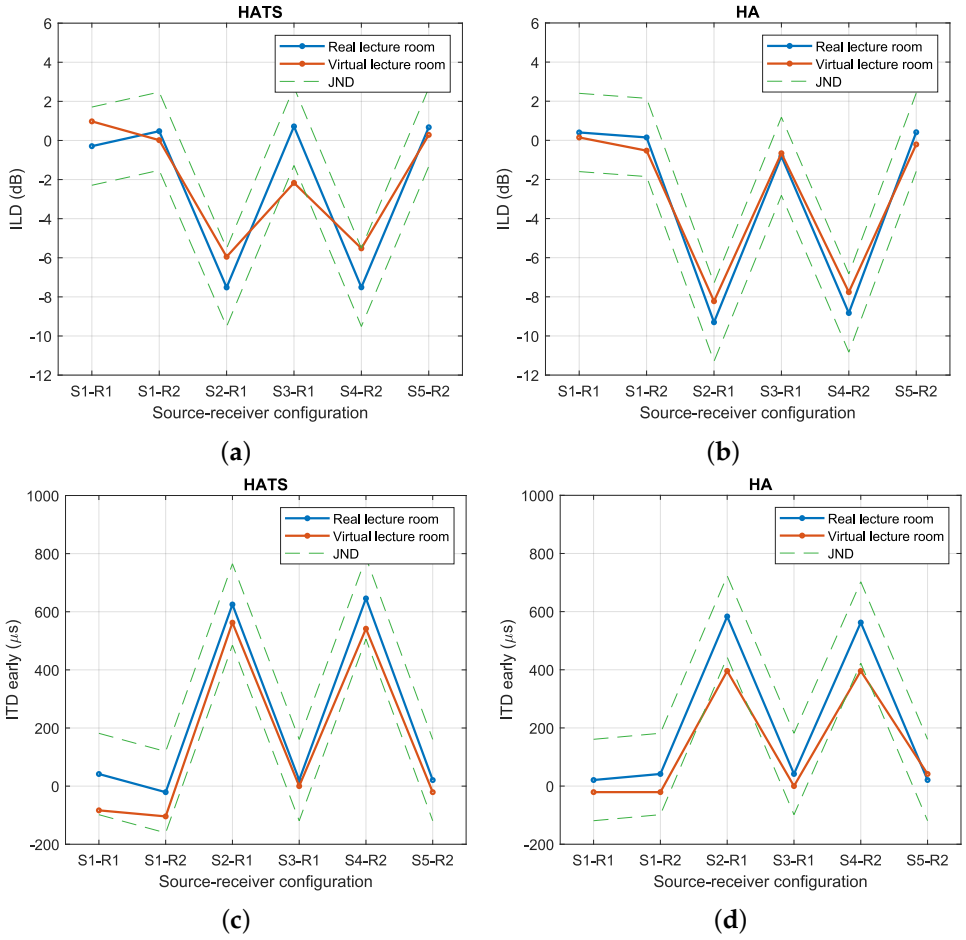
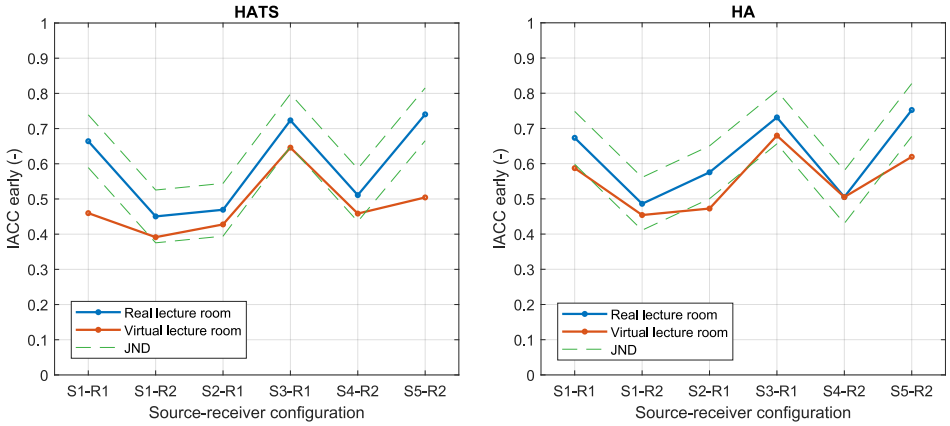


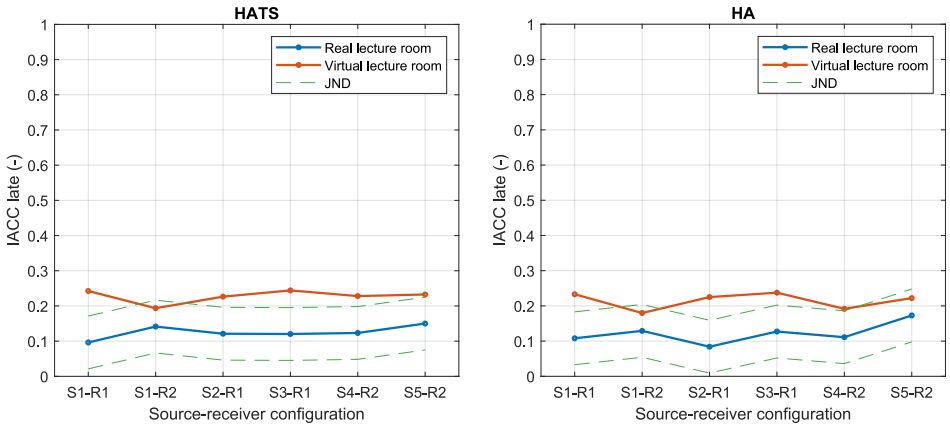
Figure 14. ILD values, band-passed from 1 to 9 kHz, for the real and the virtual lecture room and JND limits as the source–receiver configuration varies, for measurements with (**a**) HATS; (**b**) HAs. Early ITD values, band-passed from 100 to 1400 kHz, for the real and the virtual lecture room and JND limits as the source–receiver configuration varies, for measurements with (**c**) HATS; (**d**) HAs.

Finally, Figures 17 and 18 present the comparison between real and virtual lecture rooms for different off-center positions of the HATS within the ASL, specifically, at distances of 0, 10, and 20 cm, based on ILD, early ITD, and early and late IACC values. The results obtained from the intra-lab validation are confirmed, demonstrating a consistent decline in binaural parameters as the off-center distance increases. However, as expected, for all parameters except the late IACC, the values fall within the JND for receiver–source configurations with greater angular separation in the horizontal plane. This occurs because the impact of off-centering is more pronounced for sources on the median plane, as the off-centering is implemented toward the left direction.



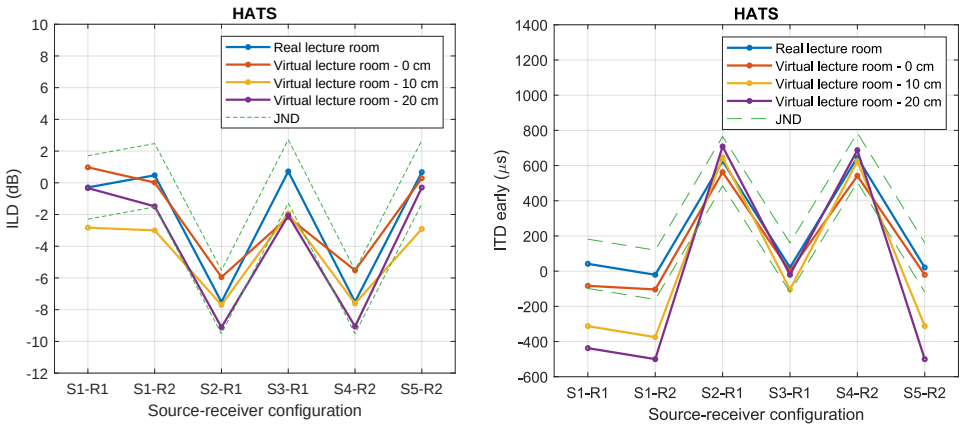
(**a**) (**b**)

Figure 15. Early IACC values averaged across the 500 to 2000 Hz octave bands and JND limits for the real and virtual lecture rooms as the source–receiver configuration varies, for measurements with (**a**) the HATS; (**b**) the HAs.



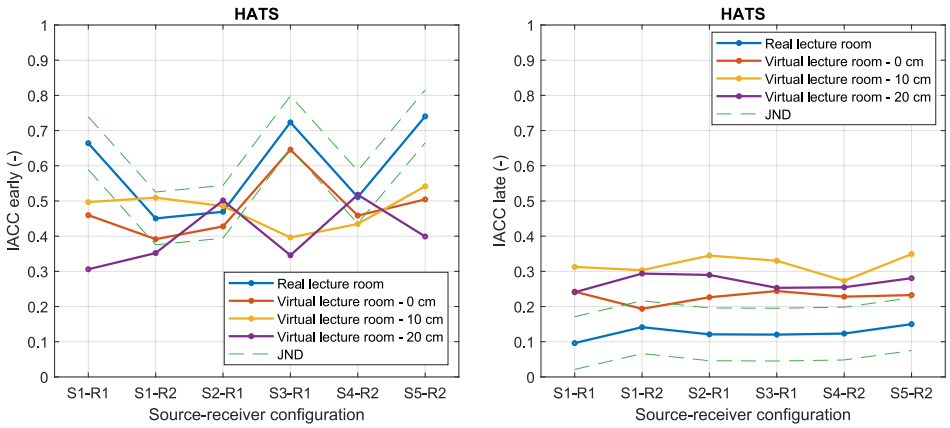
(**a**) (**b**)

Figure 16. Late IACC values averaged across the 500 to 2000 Hz octave bands and JND limits for the real and virtual lecture rooms as the source–receiver configuration varies, for measurements with (**a**) the HATS; (**b**) the HAs.



(**a**) (**b**)

Figure 17. Values computed for the real and virtual lecture rooms for different off-center positions of the HATS within the ASL as a function of the source–receiver configuration for (**a**) ILD band-passed from 1 to 9 kHz; (**b**) early ITD band-passed from 100 to 9000 Hz.

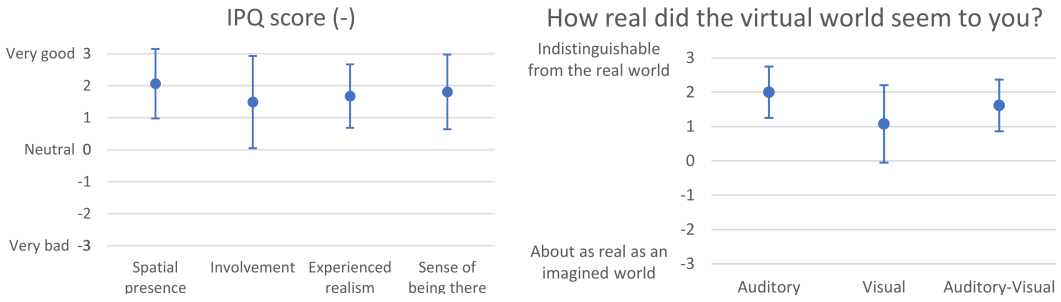


(**a**) (**b**)

Figure 18. Values computed for the real and virtual lecture rooms for different off-center positions of the HATS within the ASL as a function of the source–receiver configuration for (**a**) early IACC averaged across the 500 to 2000 Hz octave bands; (**b**) late IACC averaged across the 500 to 2000 Hz octave bands.

3.2.2. Subjective Audiovisual Evaluation

Figure 19a,b show the average and SD scores for the overall adaptation of the IPQ, divided by subscale, and for the question related to the experienced realism in terms of auditory, visual, and audiovisual components, respectively.



(**a**) (**b**)

Figure 19. (**a**) Average and SD scores for the adaptation of the IPQ divided in spatial presence, involvement, experienced realism, and sense of being there subscales. (**b**) Average and SD scores for the experienced realism question divided in auditory, visual, and auditory–visual components.

For all subscales, both the averages and SDs exceed a score of 0, which was set as the threshold above which the subjects' responses could be considered positive regarding the question: 'Can subjects envision themselves as being in the real lecture room during the virtual lecture room reproduction?'. In particular, the averages for spatial presence, experienced realism, and involvement scored 2.1, 1.7, and 1.5, respectively. In line with these scores, the single-question subscale—sense of being there—scored as the average among the other subscales, further confirming the validity of the question, which was correctly understood as the overall sense of being present in the scene that encompasses all three other subscales. Focusing on the experienced realism, the individual contributions of the auditory, visual, and audiovisual components scored 2.0, 1.1, and 1.6, respectively, highlighting a discrepancy between the perceived quality of the VAE and VVE. The question regarding the audiovisual component scored as the average between the other two components, confirming that it was correctly understood as representing overall perceived realism. Reasons for the lower score attributed to the experienced visual realism compared to the auditory realism can be found in the open comments collected from the subjects. Half of the

participants spontaneously reported that the audio felt very realistic. However, the more experienced subjects identified issues related to the 360° image reconstruction, particularly due to stitching. Additionally, a quarter of the subjects reported that the inability to read the writings on the blackboard, due to the video resolution, contributed to lowering the score assigned to the visual realism.

4. Discussion and Conclusions

Enhancing the acoustical quality in learning environments is crucial, especially for HA users. Understanding how occupants' acoustic perceptions vary within the same environment and identifying the factors that cause these differences is key to designing human-centered spaces that promote well-being. Recent advancements in VR techniques offer exciting new possibilities for studying subjective perception, particularly when in-field evaluations cannot be performed. However, most VR systems are expensive and challenging to implement in typical indoor settings, such as architectural design studios. This work, therefore, focused on validating a VR methodology that can be widely applied for subjective evaluations of room acoustical quality and comfort, particularly in educational spaces, by examining the virtual reproduction accuracy against a real reverberant lecture room. The proposed VR system is based on a 16-speaker spherical array, which renders the VAE through the well-established 3OA method and is synchronized with a VR headset. The aim was to provide a low-budget solution that could be easily installed in common non-anechoic indoor environments, where space and budget for acoustical treatment are limited. The validation process was divided into two main strands: intra-lab validation and inter-lab validation.

4.1. Intra-Lab Validation

The outcomes from the intra-lab validation highlighted that, for NH subjects, the values of the predictors of human sound localization, namely, ILD and ITD, computed for VSSs can be considered perceptually equal to those obtained for RSSs at the 0 cm off-center position from the sweet spot. This indicates that the basic and typical usage condition of the VR system ensures accurate VAE reproduction. However, VSS results begin to slightly deviate from the JND limit at higher off-center distances. A similar behavior is observed in the case of HA microphone usage. Referring to [18], which conducted a similar study with HA usage, significantly higher ΔILD values were found for a 3OA reproduction system using energy-preserving decoding and basic weighting at all off-center distances. Specifically, the average ILD errors for 0, 10, and 20 cm displacements were 4.1, 4.8, and 3.7 dB, respectively, compared to 1.5, 1.8, and 1.4 dB achieved by the system proposed in this paper, which still fall within the 2 dB JND. This confirms that by simply choosing an appropriate decoding strategy, good results can be achieved even with a low ambisonics order. Moreover, the results obtained in this study are comparable to those achieved with the best-performing rendering method, namely, VBAP on a 32-speaker array, in [18], which yielded ΔILD values of 1.7, 1.8, and 1.8 dB. However, it should be noted that Simon et al.'s study included a higher number of sound source locations—56 compared to 16. Concerning ITD, similar results to those in [18] are found at the 0 cm off-center position, but different results are observed at other off-center locations. Specifically, in [18], ITD for VSSs deviates from the desired behavior as the off-center distance increases, more than twice as much compared to the system in this paper. In the worst-case scenario—at 0° azimuth, where the JND is 20 μs—the ITD error exceeds 700 μs compared to 300 μs in our system. On average, the ITD errors obtained in this work are comparable to those achieved by [18] using VBAP and MDAP with 32 speakers. Nonetheless, because ILD and ITD errors exceed the JND for some source locations, in the case of HA and without HA for locations outside the sweet spot region, the proposed system might still lead to localization errors, which would need to be confirmed by subjects' perceptual evaluations.

4.2. Inter-Lab Validation

4.2.1. Objective Acoustical Evaluation

The overall analysis of standard monaural room acoustical parameters shows no perceptible differences between the real and virtual rooms, with the virtual room values falling within the JND range from 250 Hz to 4000 Hz, which is crucial for speech. However, at the 125 Hz octave band, differences of 0.14 s in EDT and 2.2 dB in C_{50} were observed, likely due to poor reconstruction of direct sound and early reflections. This can be attributed to a room mode around 105 Hz, confirmed by a DRR of −6 dB at the 125 Hz octave band, as discussed in Section 2.1. Additionally, a 1.5 dB difference in C_{50} was noted at the 8000 Hz octave band. Nevertheless, other studies have considered more relaxed JND limits. Yang W. and Hodgson W. [32] suggested that a JND equal to 10% for EDT could be considered as an indicator of the minimum practically significant difference. Similarly, Bradley [90] proposed that a JND value of 3 dB for C_{50} is a more reasonable estimate that aligns with the minimum clarity differences detectable in everyday listening situations. It follows that the found differences may be still imperceptible, indicating high reproduction fidelity across the 90 Hz to 9000 Hz range. This conclusion is further supported by the location-dependent analysis of speech-weighted parameters, where no perceptible differences were found. This result is significant for validating predicted speech intelligibility, as C_{50} has proven to strongly correlate with speech intelligibility in reverberant environments, even in the presence of background noise [32]. Additionally, the absence of significant differences on the DRR, a predictor of perceived sound source distance [70], further validates the system's accuracy.

The analysis of ILD with varying source–receiver configurations revealed only a minor difference between the real and virtual lecture rooms, exceeding the JND by less than 1 dB only when the sound source was behind the receiver. This may be due to ambisonics artifacts at high frequencies or slight rotational errors in HATS positioning. No perceptible differences were found for early ITD, indicating an overall accurate reconstruction of binaural cues essential for sound localization from 100 to 9000 Hz for NH, even in reverberant environments. This finding generally applies when HAs are used as well, with no significant ILD errors and only slight early ITD differences found for sources with larger angular separations on the horizontal plane.

Other spatial features, such as perceived sound source width and envelopment, are less accurately reproduced. The virtual lecture room fails to replicate the real lecture room's early and late IACC values, with differences within 0.15, for some source–receiver configurations for both HATS and HA measurements. The virtual room IACC values are lower than desired, suggesting the perception of a wider and more diffused sound, while late IACC errors suggest a reduced perceived sound envelopment. However, considering the JND values for early (0.4) and late (0.6) IACC found in [89] for guitar sounds in a reverberant environment, all IACC values measured in the virtual lecture room remain perceptually equal to the ones measured in the real lecture room. This suggests a perceptually valid reconstruction of both sound source width and envelopment. Moreover, also in [34], IACC errors of about 0.2 are considered as being within the JND for the IACC.

The analysis of the virtual lecture room accuracy at off-center positions from the sweet spot revealed the same trend observed during the intra-lab validation, based on variations in binaural parameters. For radially positioned sources, early ITD and IACC errors increase with greater off-center distances, while ILD and late IACC errors at 10 cm off-center exceed those at 20 cm. Nevertheless, IACC errors remain within the JND values reported in [89]. These findings emphasize the need for conducting subjective localization tests as a benchmark for assessing localization errors caused by slight listener's movements outside the sweet spot.

4.2.2. Subjective Audiovisual Evaluation

The questions from the adapted IPQ, collected from students who were used to attending the real lecture room, scored above 0 for all subscales, with an average score

of 1.8—labeled as 'almost very good'. This indicates that the subjects had a positive experience that closely resembled real life; specifically, the participants felt as though they were physically present in the real lecture room. Additionally, they were engaged and devoted their attention to the lecture scene. Then, concerning the contributions of the auditory and visual components to the experienced realism, the subjects rated their auditory experience at the very positive end of the scale but rated the visual experience somewhat lower. This was further confirmed by the open interviews, in which half of the subjects spontaneously appreciated the perceived auditory realism while highlighting some video post-processing and resolution issues that negatively affected the visual sense of realism. Thus, there is still room for improvement in the visual component of the VR system, which would not only enhance realism but also contribute to a stronger overall sense of presence. The literature shows that integrating a visual virtual environment (VVE) is essential for enhancing immersion and reinforcing auditory illusions. Visual cues improve realism and encourage movements that mimic real-life listening behavior, influencing sound source localization and speech intelligibility, particularly with lip-reading cues. However, misalignment or varying quality between visual and auditory elements can cause confusion or cognitive dissonance for the listener [91]. This dissonance can hinder accurate sound source perception and speech comprehension, undermining the auditory illusion and overall room acoustic quality. In this context, the inferior visual reproduction in the proposed VR system may have diminished the sense of presence. Therefore, enhancing visual quality is likely to improve subjective outcomes.

4.3. Limitations and Future Perspectives

Although the validation in this work aimed to closely align the measurements with a real listener's experience in both real and VEs thereby providing a true assessment of subjective perception, some limitations related to the measurement equipment and the need for further evaluations must be acknowledged. The current study utilized an HATS optimized for soundscape measurements; however, models with more human-like features, such as more defined and pronounced pinnae and a more realistic ear canal, could be employed in future research developments. These enhancements would be particularly valuable for re-evaluating binaural parameters that are sensitive to high frequencies, such as ILD and, to some extent, IACC, to determine whether the identified errors exceeding the JND still persist. Furthermore, incorporating spectral differences analysis could provide deeper insights into the perceived elevation of VSSs. To this end, performing sound localization tests with human subjects may be a valuable approach to further validate the outcomes of the objective predictor metrics, ultimately leading to a more definitive assessment of the VR system's performance in accurately reproducing the spatial features of sound. Similarly, speech intelligibility tests directly involving subjects, such as SRT measurements, should be conducted to compare real and virtual rooms, thereby gaining robust confirmation of the results derived from the speech intelligibility predictor metrics used in this paper. This would provide definitive evidence of the system's effectiveness in evaluating educational spaces where accurate speech intelligibility is crucial. Moreover, all measurements performed with HAs refer to the pre-beamformer case. Therefore, valuable insights into the validity of this VR system for HA users could be gained by repeating the validation procedure with different beamformer algorithms. A key consideration is that our goal is to develop a widely applicable methodology for assessing acoustic quality and comfort in educational environments. The VR system must be inclusive, allowing hearing-impaired (HI) individuals and hearing aid (HA) users to evaluate acoustic adequacy, as good quality is essential for compensating communicative deficits. Although this presents challenges, future validation should be user-oriented. The variety of hearing impairments and devices complicates universal validation, requiring a multi-step approach targeting different subject categories. Following studies on speech intelligibility with normal-hearing (NH) individuals, we will focus on tests with various HI groups and HA users. Moreover, regarding other limitations and potential improvements for the VR system

that could further enhance the overall sense of presence experienced by subjects during the audiovisual virtual experience, it was found that the video resolution and post-processing procedures significantly affected the sense of being in the real lecture room. However, these issues can be easily addressed in the future by implementing more careful and advanced video post-processing techniques and utilizing higher 360° video resolutions. The current VAE reproduction system uses two Genelec 8351A monitors for a narrow frequency range (30 to 90 Hz). Replacing them with a more cost-effective subwoofer, like the Genelec 7050, could optimize efficiency without sacrificing audio quality. Furthermore, while both the acoustical and visual components of the reproduced immersive scenes are based on in-field measurements, to exclude during the initial validation phase potential errors not strictly due to the VR system but rather due to the use of AV simulations, future implementations could include 3D models of the environments coupled with acoustic simulations and real-time subject navigation within them. This future extension will allow for the perceptual validation of new buildings during the design phase by integrating acoustic simulation software into the audiovisual chain and developing a custom application in Unreal Engine, thereby improving visual realism and interaction capabilities. To promote broader use of the proposed VR-based methodology, we plan to develop a user-friendly application that automates the input of new environments and simplifies playback through the VR system, along with available perceptual tests. Additionally, we will provide detailed step-by-step installation guidelines for the system using the proposed commercial components, including an automated audio tuning procedure, alongside the free software framework.

4.4. Final Remarks

The proposed VR system successfully captures the key acoustical properties of the real lecture room, both in terms of standard room acoustical parameters and, more importantly, in terms of speech-weighted parameters related to perceived speech intelligibility. It accurately reproduces sound localization at the sweet spot in both acoustically treated and reverberant environments, with ILD and ITD values for virtual sources closely matching those of real sources, even when HA microphones are used. However, outside the sweet spot, ILD and ITD errors for some sound source locations exceed the JND, potentially leading to localization errors that require further subjective validations. The perceived spatial impression of sound has also been satisfactorily reproduced by the VR system, with IACC errors falling within most accepted JND limits. The subjective audiovisual evaluation of the VR system revealed that students felt a strong sense of presence in the lecture room, indicating an audiovisual experience that closely resembled real life. However, weaknesses in the visual component of the VR system were identified, which can be easily addressed in the future to enhance the overall sense of presence during the virtual experience. It follows that the proposed VR system is well suited to be adopted as a methodology for predicting the perceived acoustical quality of rooms, particularly educational spaces, by facilitating the execution of subjective tests. So far, in fact, tests with students in field have been carried out in a limited number of studies but require a great effort, as they are based on the use of speech intelligibility tests that have been validated and optimized even for vulnerable categories [e.g., [92]], but are especially time consuming. While it may not serve as a precision instrument, it performs well enough to provide a general impression, as evidenced by the validation results. This makes it suitable for its intended application—assessing perceived acoustical comfort of educational buildings—especially considering its low cost and the limited effort required for implementation compared to other existing systems. Beyond learning spaces, the same VR system could be applied to evaluate other environments, such as conference rooms, coworking spaces, and open-plan offices, as well as theaters and concert halls. Moreover, the system's purpose could even be reversed. It could be used in clinical practice to facilitate a more efficient assessment of patients' hearing loss and enable more effective HA fittings by testing them in patient-tailored auditory scenarios that emulate real-life HA usage conditions [25,74,93]. Lastly, the VR system could also be

extended to more engaging fields, such as gaming or entertainment, where immersive and realistic acoustics enhance the user experience.

Author Contributions: Conceptualization, A.G., G.E.P. and A.A.; data curation, A.G.; formal analysis, A.G.; investigation, A.G.; methodology, A.G., L.S., G.E.P. and A.A.; project administration, A.A., A.G. and F.R.; resources, A.A.; software, A.G. and F.R.; supervision, A.A., G.E.P., F.R. and L.S.; validation, L.S., F.R. and A.A.; visualization, A.G.; writing—original draft preparation, A.G. and R.G.R.; writing—review and editing, A.A., F.R., G.E.P. and L.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Institutional Review Board (or Ethics Committee) of Politecnico di Torino (100993/2023).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Acknowledgments: We would like to express our gratitude to Riccardo Caradonna, Andrea Galletto, Andrea Gerbotto, Riccardo Lacqua, and Lorenzo Lavagna for their assistance during the in-field measurements of the lecture room. Special thanks go to Università degli studi di Parma for providing the Eigenmike em64, to HEAD acoustics for supplying the BSU HATS, and to the Carl von Ossietzky University of Oldenburg for the HAs. We also extend our appreciation to Marco Pagliano Sasso, Federico Palumbo, Nicolò Somà, and Filippo Trematerra for their help with the subjective audiovisual tests.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

3D	Three dimensional
3OA	Third-order ambisonics
5OA	Fifth-order ambisonics
AllRAD	All-Round Ambisonic Decoding
ASL	Audio Space Lab
AV	Audiovisual
BRIR	Binaural room impulse response
C_{50}, C_{80}	Clarity
CAVE	Cave Automatic Virtual Environment
CTC	Cross-talk cancellation
DAW	Digital audio workstation
DBAP	Distance-based amplitude panning
DRR	Direct-to-reverberant ratio
EDT	Early decay time
HA	Hearing aid
HATS	Head and Torso Simulator
HI	Hearing impaired
HOA	High-order ambisonics
HRTF	Head-related transfer function
HVAC	Heating, ventilation, and air conditioning
IACC	Interaural cross-correlation

IIR	Infinite impulse response
ILD	Interaural level difference
IPQ	Igroup Presence Questionnaire
ITD	Interaural time difference
JND	Just-noticeable difference
MDAP	Multiple-direction amplitude panning
NH	Normal hearing
NSP	Nearest-speaker panning
OSC	Open Sound Control
RIR	Room impulse response
RSS	Real sound source
SD	Standard deviation
SMA	Spherical microphone array
SLM	Sound level meter
SRT	Speech reception threshold
STI	Speech Transmission Index
STIPA	Speech Transmission Index for Public Addresses
T_{60}, T_{30}, T_{20}	Reverberation time
VAE	Virtual acoustic environment
VBAP	Vector base amplitude panning
VE	Virtual environment
VSS	Virtual sound source
VVE	Virtual visual environment
VR	Virtual reality
WFS	Wave field synthesis
WOW	Window on world

Appendix A

Table A1. Questionnaire used for the subjective audiovisual evaluation of the VR system adapted from the IPQ.

Items	Subscales	English Question	English Anchors	Italian Question	Italian Anchors
1	Sense of being there	In the computer generated world I had a sense of "being there"	Not at all—Very much	Nel mondo virtuale, avevo la sensazione di "essere lì".	Per niente—Moltissimo
2	Spatial presence	Somehow I felt that the virtual world surrounded me	Fully disagree—Fully agree	In qualche modo ho avvertito che il mondo virtuale mi circondasse.	Completamente in disaccordo—Completamente d'accordo
3	Spatial presence	I felt like I was just perceiving pictures	Fully disagree—Fully agree	Mi sembrava come se stessi solo percependo delle immagini.	Completamente in disaccordo—Completamente d'accordo
6	Spatial presence	I felt present in the virtual space.	Fully disagree—Fully agree	Mi sono sentito come se fossi stato realmente presente nell'ambiente virtualizzato.	Completamente in disaccordo—Completamente d'accordo
7	Involvement	How aware were you of the real world surrounding while navigating in the virtual world? (i.e. sounds, room temperature, other people, etc.)?	Extremely aware—Moderately aware—Not aware at all	Quanto eri consapevole del mondo reale circostante mentre navigavi nel mondo virtuale? (ad esempio, suoni, temperatura dell'ambiente, altre persone, ecc.)?	Completamente consapevole—Moderatamente consapevole—Completamente inconsapevole
8	Involvement	I was not aware of my real environment.	Fully disagree—Fully agree	Non ero consapevole del mio ambiente reale	Completamente in disaccordo—Completamente d'accordo
9	Involvement	I still paid attention to the real environment.	Fully disagree—Fully agree	Continuavo ancora a prestare attenzione all'ambiente reale.	Completamente d'accordo—Completamente in disaccordo
10	Involvement	I was completely captivated by the virtual world.	Fully disagree—Fully agree	Ero completamente affascinato dal mondo virtuale.	Completamente in disaccordo—Completamente d'accordo
11	Experienced Realism	How real did the virtual world seem to you?	Completely real—Not real at all	Quanto reale ti è sembrato il mondo virtuale?	Per niente reale—Moderatamente reale—Completamente reale
12	Experienced Realism	How much did your experience in the VE seem consistent with your real world experience?	Not consistent—Moderately consistent—Very consistent	In che misura la tua esperienza nell'ambiente virtuale sembrava coerente con la tua esperienza del mondo reale?	Non coerente—Moderatamente coerente—Molto coerente
13	Experienced Realism	How real did the virtual auditory-visual world seem to you?	About as real as an imagined world—Indistinguishable from the real world	Quanto reale ti è sembrato il mondo virtuale audiovisivo?	All'incirca reale quanto un mondo immaginato—Indistinguibile dal mondo reale
13a*	Experienced Acoustical Realism	How real did the virtual acoustical world seem to you?	About as real as an imagined world—Indistinguishable from the real world	Quanto reale ti è sembrato il mondo virtuale acustico?	All'incirca reale quanto un mondo immaginato—Indistinguibile dal mondo reale
13b*	Experienced Visual Realism	How real did the virtual visual world seem to you?	About as real as an imagined world—Indistinguishable from the real world	Quanto reale ti è sembrato il mondo virtuale visivo?	All'incirca reale quanto un mondo immaginato—Indistinguibile dal mondo reale

References

1. Astolfi, A. Premises for Effective Teaching and Learning: State of the Art, New Outcomes and Perspectives of Classroom Acoustics. *Int. J. Acoust. Vib* **2023**, *28*, 86–97. [CrossRef]
2. Peelle, J.E. Listening effort: How the cognitive consequences of acoustic challenge are reflected in brain and behavior. *Ear Hear.* **2018**, *39*, 204–214. [CrossRef] [PubMed]
3. Puglisi, G.E.; Warzybok, A.; Astolfi, A.; Kollmeier, B. Effect of reverberation and noise type on speech intelligibility in real complex acoustic scenarios. *Build. Environ.* **2021**, *204*, 108137. [CrossRef]
4. Puglisi, G.E.; Di Iulio, M.; Bottalico, P.; Murgia, S.; Consolino, P.; Bisetti, M.S.; Pittà, G.; Shtrepi, L.; Astolfi, A. Challenges for Children with Cochlear Implants in Everyday Listening Scenarios: The Competitive Effect of Noise and Face Masks on Speech Intelligibility. *Appl. Sci.* **2023**, *13*, 8715. [CrossRef]
5. Seeber, B.U.; Kerber, S.; Hafter, E.R. A system to simulate and reproduce audio–visual environments for spatial hearing research. *Hear. Res.* **2010**, *260*, 1–10. [CrossRef] [PubMed]
6. Pausch, F.; Aspöck, L.; Vorländer, M.; Fels, J. An extended binaural real-time auralization system with an interface to research hearing aids for experiments on subjects with hearing loss. *Trends Hear.* **2018**, *22*, 2331216518800871. [CrossRef]
7. Brimijoin, W.O.; Boyd, A.W.; Akeroyd, M.A. The contribution of head movement to the externalization and internalization of sounds. *PLoS ONE* **2013**, *8*, e83068. [CrossRef]
8. Simon, L.S.; Wuethrich, H.; Dillier, N. *Comparison of Higher-Order Ambisonics, Vector- and Distance-Based Amplitude Panning Using a Hearing Device Beamformer*; University of Zurich: Zürich, Switzerland, 2017.
9. Nykänen, A.; Zedigh, A.; Mohlin, P. Effects on localization performance from moving the sources in binaural reproductions. In Proceedings of the International Congress and Exposition on Noise Control Engineering, Innsbruck, Austria, 15–18 September 2013; ÖAL Österreichischer Arbeitsring für Lärmbekämpfung: Wien, Austria, 2013; pp. 3193–3201.
10. Völk, F.; Heinemann, F.; Fastl, H. Externalization in binaural synthesis: Effects of recording environment and measurement procedure. In Proceedings of the Acoustics 08, Paris, France, 30 June–4 July 2008; pp. 6419–6424.
11. Berkhout, A.J.; de Vries, D.; Vogel, P. Acoustic control by wave field synthesis. *J. Acoust. Soc. Am.* **1993**, *93*, 2764–2778. [CrossRef]
12. Grimm, G.; Ewert, S.; Hohmann, V. Evaluation of spatial audio reproduction schemes for application in hearing aid research. *Acta Acust. United Acust.* **2015**, *101*, 842–854. [CrossRef]
13. Pulkki, V. Virtual sound source positioning using vector base amplitude panning. *J. Audio Eng. Soc.* **1997**, *45*, 456–466.
14. Kostadinov, D.; Reiss, J.D.; Mladenov, V.M. Evaluation of Distance Based Amplitude panning for spatial audio. In Proceedings of the ICASSP, Dallas, TX, USA, 14–19 March 2010; pp. 285–288.
15. Frank, M. Localization using different amplitude-panning methods in the frontal horizontal plane. In Proceedings of the EAA Joint Symposium on Auralization and Ambisonics, Berlin, Germany, 3–5 April 2014.
16. Oreinos, C.; Buchholz, J.M. Objective analysis of ambisonics for hearing aid applications: Effect of listener's head, room reverberation, and directional microphones. *J. Acoust. Soc. Am.* **2015**, *137*, 3447–3465. [CrossRef] [PubMed]
17. Gerken, M.; Hohmann, V.; Grimm, G. Comparison of 2D and 3D Multichannel Audio Rendering Methods for Hearing Research Applications using Technical and Perceptual Measures. *Acta Acust.* **2024**, *8*, 17. [CrossRef]
18. Simon, L.S.; Dillier, N.; Wüthrich, H. Comparison of 3D audio reproduction methods using hearing devices. *J. Audio Eng. Soc.* **2021**, *68*, 899–909. [CrossRef]
19. Llorach, G.; Grimm, G.; Hendrikse, M.M.; Hohmann, V. Towards realistic immersive audiovisual simulations for hearing research: Capture, virtual scenes and reproduction. In Proceedings of the 2018 Workshop on Audio-Visual Scene Understanding for Immersive Multimedia, Seoul, Republic of Korea, 22–26 October 2018; pp. 33–40.
20. Pausch, F.; Behler, G.; Fels, J. *SCaLAr—A Surrounding Spherical Cap Loudspeaker Array for Flexible Generation and Evaluation of Virtual Acoustic Environments*; EDP Sciences: Les Ulis, France, 2020.
21. Favrot, S.; Buchholz, J.M. LoRA: A loudspeaker-based room auralization system. *Acta Acust. United Acust.* **2010**, *96*, 364–375. [CrossRef]
22. Nachbar, C.; Zotter, F.; Deleflie, E.; Sontacchi, A. Ambix-a suggested ambisonics format. In Proceedings of the Ambisonics Symposium, Lexington, KT, USA, 2–3 July 2011; Volume 2011.
23. Frank, M. Phantom Sources Using Multiple Loudspeakers in the Horizontal Plane. Ph.D. Thesis, Institute of Electronic Music and Acoustics, University of Music and Performing Arts Graz, Graz, Austria, 2013.
24. Frank, M. How to make Ambisonics sound good. In Proceedings of the Forum Acusticum, Krakow, Poland, 7–12 September 2014.
25. Cubick, J.; Dau, T. Validation of a virtual sound environment system for testing hearing aids. *Acta Acust. United Acust.* **2016**, *102*, 547–557. [CrossRef]
26. Wierstorf, H.; Raake, A.; Spors, S. Assessing localization accuracy in sound field synthesis. *J. Acoust. Soc. Am.* **2017**, *141*, 1111–1119. [CrossRef]
27. Zotter, F.; Frank, M. All-round ambisonic panning and decoding. *J. Audio Eng. Soc.* **2012**, *60*, 807–820.
28. Frank, M.; Marentakis, G.; Sontacchi, A. A simple technical measure for the perceived source width. In Proceedings of the Fortschritte der Akustik, DAGA, Düsseldorf, Germany, 21–24 March 2011.
29. Bertet, S.; Daniel, J.; Parizet, E.; Warusfel, O. Investigation on localisation accuracy for first and higher order ambisonics reproduced sound sources. *Acta Acust. United Acust.* **2013**, *99*, 642–657. [CrossRef]

30. *EN ISO 3382-1*; Acoustics—Measurement of Room Acoustic Parameters—Part 2: Performance Spaces. International Organization for Standardization: Genève, Switzerland, 2009.
31. Fargeot, S.; Vidal, A.; Aramaki, M.; Kronland-Martinet, R. Perceptual evaluation of an ambisonic auralization system of measured 3D acoustics. *Acta Acust.* **2023**, *7*, 56. [CrossRef]
32. Yang, W.; Hodgson, M. Validation of the auralization technique: Comparative speech-intelligibility tests in real and virtual classrooms. *Acta Acust. United Acust.* **2007**, *93*, 991–999.
33. Ahrens, A.; Marschall, M.; Dau, T. Evaluating the auralization of a small room in a virtual sound environment using objective room acoustic measures. In Proceedings of the 5th Joint Meeting of the Acoustical Society of America and Acoustical Society of Japan, Honolulu, HI, USA, 28 November–2 December 2016.
34. Hládek, L.; Ewert, S.D.; Seeber, B.U. Communication conditions in virtual acoustic scenes in an underground station. In Proceedings of the IEEE 2021 Immersive and 3D Audio: From Architecture to Automotive (I3DA), Bologna, Italy, 8–10 September 2021; pp. 1–8.
35. Álvarez-Morales, L.; Galindo, M.; Girón, S.; Zamarreño, T.; Cibrián, R. Acoustic characterisation by using different room acoustics software tools: A comparative study. *Acta Acust. United Acust.* **2016**, *102*, 578–591. [CrossRef]
36. Ahrens, A.; Lund, K.D.; Marschall, M.; Dau, T. Sound source localization with varying amount of visual information in virtual reality. *PLoS ONE* **2019**, *14*, e0214603. [CrossRef]
37. *BS EN IEC 60268-16:2020*; Sound System Equipment—Part 16: Objective Rating of Speech Intelligibility by Speech Transmission Index. BSI Standards Publication: London, UK, 2020.
38. Favrot, S.; Buchholz, J.M. Distance perception in loudspeaker-based room auralization. In Proceedings of the 127th Audio Engineering Society Convention, New York, NY, USA, 9–10 October 2009.
39. Neidhardt, A.; Schneiderwind, C.; Klein, F. Perceptual matching of room acoustics for auditory augmented reality in small rooms-literature review and theoretical framework. *Trends Hear.* **2022**, *26*, 23312165221092919. [CrossRef]
40. Hendrikse, M.M.; Llorach, G.; Grimm, G.; Hohmann, V. Influence of visual cues on head and eye movements during listening tasks in multi-talker audiovisual environments with animated characters. *Speech Commun.* **2018**, *101*, 70–84. [CrossRef]
41. Grimm, G.; Hendrikse, M.M.E.; Hohmann, V. Review of Self-Motion in the Context of Hearing and Hearing Device Research. *Ear Hear.* **2020**, *41*, 48S–55S. [CrossRef]
42. Hendrikse, M.M.; Eichler, T.; Hohmann, V.; Grimm, G. Self-motion with hearing impairment and (directional) hearing aids. *Trends Hear.* **2022**, *26*, 23312165221078707. [CrossRef] [PubMed]
43. Guastamacchia, A.; Riente, F.; Shtrepi, L.; Puglisi, G.E.; Pellerey, F.; Astolfi, A. Speech intelligibility in reverberation based on audio-visual scenes recordings reproduced in a 3D virtual environment. *Build. Environ.* **2024**, *258*, 111554. [CrossRef]
44. Grant, K.W. The effect of speechreading on masked detection thresholds for filtered speech. *J. Acoust. Soc. Am.* **2001**, *109*, 2272–2275. [CrossRef]
45. MacLeod, A.; Summerfield, Q. Quantifying the contribution of vision to speech perception in noise. *Br. J. Audiol.* **1987**, *21*, 131–141. [CrossRef]
46. Feng, Y.; Duives, D.C.; Hoogendoorn, S.P. Wayfinding behaviour in a multi-level building: A comparative study of HMD VR and Desktop VR. *Adv. Eng. Inform.* **2022**, *51*, 101475. [CrossRef]
47. Kwon, C. Verification of the possibility and effectiveness of experiential learning using HMD-based immersive VR technologies. *Virtual Real.* **2019**, *23*, 101–118. [CrossRef]
48. Bolaños, J.G.; Pulkki, V. Immersive audiovisual environment with 3D audio playback. In Proceedings of the 132nd Audio Engineering Society Convention, Budapest, Hungary, 26–29 April 2012.
49. Cruz-Neira, C.; Sandin, D.J.; DeFanti, T.A.; Kenyon, R.V.; Hart, J.C. The CAVE: Audio visual experience automatic virtual environment. *Commun. ACM* **1992**, *35*, 64–73. [CrossRef]
50. Muhanna, M.A. Virtual reality and the CAVE: Taxonomy, interaction challenges and research directions. *J. King Saud Univ.-Comput. Inf. Sci.* **2015**, *27*, 344–361. [CrossRef]
51. Shibata, T. Head mounted display. *Displays* **2002**, *23*, 57–64. [CrossRef]
52. Rolland, J.P.; Cakmakci, O. The past, present, and future of head-mounted display designs. *Proc. SPIE* **2005**, *5638*, 368–377.
53. Capron, D.W.; Norr, A.M.; Albanese, B.J.; Schmidt, N.B. Fear reactivity to cognitive dyscontrol via novel head-mounted display perceptual illusion exercises. *J. Affect. Disord.* **2017**, *217*, 138–143. [CrossRef]
54. van Heugten-van der Kloet, D.; Cosgrave, J.; van Rheede, J.; Hicks, S. Out-of-body experience in virtual reality induces acute dissociation. *Psychol. Conscious. Theory Res. Pract.* **2018**, *5*, 346. [CrossRef]
55. Ramaseri Chandra, A.N.; El Jamiy, F.; Reza, H. A systematic survey on cybersickness in virtual environments. *Computers* **2022**, *11*, 51. [CrossRef]
56. Thorp, S.; Sævild Ree, A.; Grassini, S. Temporal development of sense of presence and cybersickness during an immersive vr experience. *Multimodal Technol. Interact.* **2022**, *6*, 31. [CrossRef]
57. Privitera, A.G.; Fontana, F.; Geronazzo, M. On the Effect of User Tracking on Perceived Source Positions in Mobile Audio Augmented Reality. In Proceedings of the 15th Biannual Conference of the Italian SIGCHI Chapter, Torino, Italy, 20–22 September 2023; pp. 1–9.
58. Marshall, L.G. An acoustics measurement program for evaluating auditoriums based on the early/late sound energy ratio. *J. Acoust. Soc. Am.* **1994**, *96*, 2251–2261. [CrossRef]

59. Campbell, C.; Nilsson, E.; Svensson, C. The same reverberation time in two identical rooms does not necessarily mean the same levels of speech clarity and sound levels when we look at impact of different ceiling and wall absorbers. *Energy Procedia* **2015**, *78*, 1635–1640. [CrossRef]
60. Puglisi, G.E.; Prato, A.; Sacco, T.; Astolfi, A. Influence of classroom acoustics on the reading speed: A case study on Italian second-graders. *J. Acoust. Soc. Am.* **2018**, *144*, EL144–EL149. [CrossRef] [PubMed]
61. Katz, B.F.; Noisternig, M. A comparative study of interaural time delay estimation methods. *J. Acoust. Soc. Am.* **2014**, *135*, 3530–3540. [CrossRef] [PubMed]
62. Loiselle, L.H.; Dorman, M.F.; Yost, W.A.; Cook, S.J.; Gifford, R.H. Using ILD or ITD cues for sound source localization and speech understanding in a complex listening environment by listeners with bilateral and with hearing-preservation cochlear implants. *J. Speech Lang. Hear. Res.* **2016**, *59*, 810–818. [CrossRef] [PubMed]
63. RECOMMENDATION ITU-R BS.1116-3—Methods for the Subjective Assessment of Small Impairments in Audio Systems. R BS. Available online: https://api.semanticscholar.org/CorpusID:140119719 (accessed on 27 August 2024).
64. Astolfi, A.; Riente, F.; Shtrepi, L.; Carullo, A.; Scopece, L.; Masoero, M. Speech quality improvement of commercial flat screen TV-sets. *IEEE Trans. Broadcast.* **2021**, *67*, 685–695. [CrossRef]
65. Spors, S.; Wierstorf, H.; Raake, A.; Melchior, F.; Frank, M.; Zotter, F. Spatial sound with loudspeakers and its perception: A review of the current state. *Proc. IEEE* **2013**, *101*, 1920–1938. [CrossRef]
66. Available online: https://audioplugins.iem.sh/website/docs/plugindescriptions/ (accessed on 24 August 2024).
67. Daniel, J.; Rault, J.B.; Polack, J.D. Ambisonics encoding of other audio formats for multiple listening conditions. In Proceedings of the 105th Audio Engineering Society Convention, San Francisco, CA, USA, 26–29 September 1998.
68. Steeneken, H.J.; Houtgast, T. Mutual dependence of the octave-band weights in predicting speech intelligibility. *Speech Commun.* **1999**, *28*, 109–123. [CrossRef]
69. Hummersone, C. Impulse Response Acoustic Information Calculator. GitHub. 2023. Available online: https://github.com/IoSR-Surrey/MatlabToolbox (accessed on 30 December 2023).
70. Zahorik, P. Direct-to-reverberant energy ratio sensitivity. *J. Acoust. Soc. Am.* **2002**, *112*, 2110–2117. [CrossRef]
71. Tucker, A.J.; Martens, W.L.; Dickens, G.; Hollier, M.P. Perception of reconstructed sound-fields: The dirty little secret. In Proceedings of the Audio Engineering Society Conference: 52nd International Conference: Sound Field Control-Engineering and Perception, Guildford, UK, 2–4 September 2013.
72. Santala, O.; Vertanen, H.; Pekonen, J.; Oksanen, J.; Pulkki, V. Effect of listening room on audio quality in Ambisonics reproduction. In Proceedings of the 126th Audio Engineering Society Convention, Munich, Germany, 7–10 May 2009.
73. Epic Games: Unreal Engine 5. Available online: www.unrealengine.com (accessed on 24 August 2024).
74. Guastamacchia, A.; Ebri, M.; Bottega, A.; Armelloni, E.; Farina, A.; Puglisi, G.E.; Riente, F.; Shtrepi, L.; Masoero, M.C.; Astolfi, A.; et al. Set up and preliminary validation of a small spatial sound reproduction system for clinical purposes. In Proceedings of the Forum Acusticum, Torino, Italy, 11–15 September 2023; pp. 4991–4998.
75. Portable Hearing Lab Headsets from BatAndCat Sound Labs. Available online: https://batandcat.com/portable-hearing-laboratory-phl.html (accessed on 24 August 2024).
76. Available online: https://git.rwth-aachen.de/ita/toolbox (accessed on 26 August 2024).
77. Middlebrooks, J.C. Sound localization. *Handb. Clin. Neurol.* **2015**, *129*, 99–116.
78. Benichoux, V.; Rébillat, M.; Brette, R. On the variation of interaural time differences with frequency. *J. Acoust. Soc. Am.* **2016**, *139*, 1810–1821. [CrossRef]
79. *EN ISO 3382-2*; Acoustics—Measurement of Room Acoustic Parameters—Part 2: Reverberation Time in Ordinary Rooms. International Organization for Standardization: Genève, Switzerland, 2008.
80. Farina, A.; Capra, A.; Chiesi, L.; Scopece, L. A spherical microphone array for synthesizing virtual directive microphones in live broadcasting and in post production. In Proceedings of the Audio Engineering Society Conference: 40th International Conference: Spatial Audio: Sense the Sound of Space, Tokyo, Japan, 8–10 October 2010.
81. Dietrich, P.; Guski, M.; Klein, J.; Müller-Trapet, M.; Pollow, M.; Scharrer, R.; Vorländer, M. Measurements and room acoustic analysis with the ITA-Toolbox for MATLAB. In Proceedings of the 40th Italian (AIA) Annual Conference on Acoustics and the 39th German Annual Conference on Acoustics (DAGA), Merano, Italy, 18–21 March 2013; p. 50.
82. Litovsky, R.Y.; Colburn, H.S.; Yost, W.A.; Guzman, S.J. The precedence effect. *J. Acoust. Soc. Am.* **1999**, *106*, 1633–1654. [CrossRef]
83. Bradley, J.S.; Soulodre, G.A. The influence of late arriving energy on spatial impression. *J. Acoust. Soc. Am.* **1995**, *97*, 2263–2271. [CrossRef]
84. Okano, T. Judgments of noticeable differences in sound fields of concert halls caused by intensity variations in early reflections. *J. Acoust. Soc. Am.* **2002**, *111*, 217–229. [CrossRef]
85. Guastamacchia, A.; Puglisi, G.; Bottega, A.; Shtrepi, L.; Riente, F.; Astolfi, A. *Influence of Stand Configurations on Ecological Validity of Audiovisual Recording Systems*; RWTH Publications: Aachen, Germany, 2023; pp. 88–91.
86. Igroup Presence Questionnaire (IPQ). Available online: http://www.igroup.org/pq/ipq/index.php (accessed on 27 August 2024).
87. Mossop, J.E.; Culling, J.F. Lateralization of large interaural delays. *J. Acoust. Soc. Am.* **1998**, *104*, 1574–1579. [CrossRef] [PubMed]
88. Hancock, K.E.; Delgutte, B. A physiologically based model of interaural time difference discrimination. *J. Neurosci.* **2004**, *24*, 7110–7117. [CrossRef] [PubMed]

89. Klockgether, S.; van de Par, S. Just noticeable differences of spatial cues in echoic and anechoic acoustical environments. *J. Acoust. Soc. Am.* **2016**, *140*, EL352–EL357. [CrossRef]
90. Bradley, J.S.; Reich, R.; Norcross, S. A just noticeable difference in C50 for speech. *Appl. Acoust.* **1999**, *58*, 99–108. [CrossRef]
91. Siddig, A.; Sun, P.W.; Parker, M.; Hines, A. Perception Deception: Audio-Visual Mismatch in Virtual Reality Using the McGurk Effect. *AICS* **2019**, *2019*, 176–187.
92. Puglisi, G.E.; di Berardino, F.; Montuschi, C.; Sellami, F.; Albera, A.; Zanetti, D.; Albera, R.; Astolfi, A.; Kollmeier, B.; Warzybok, A. Evaluation of Italian simplified matrix test for speech-recognition measurements in noise. *Audiol. Res.* **2021**, *11*, 73–88. [CrossRef]
93. Mehra, R.; Brimijoin, O.; Robinson, P.; Lunner, T. Potential of Augmented Reality Platforms to Improve Individual Hearing Aids and to Support More Ecologically Valid Research. *Ear Hear.* **2020**, *41*, 140S–146S. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.