







# Co-immersion in Audio Augmented Virtuality: The Case Study of a Static and Approximated Late Reverberation Algorithm

Davide Fantini  Giorgio Presti  Michele Geronazzo , Senior Member, IEEE, Riccardo Bona   
Alessandro Giuseppe Privitera  Federico Avanzini 

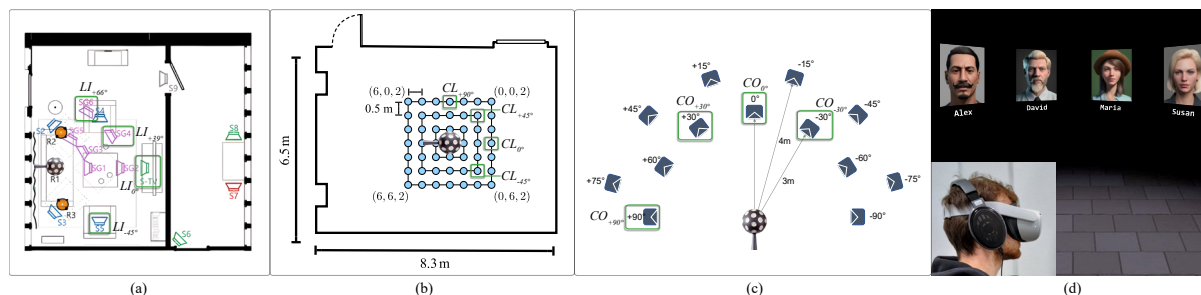


Fig. 1: The three virtual acoustic environments (VAEs) evaluated in our experiment: (a) living room [61], (b) classroom [56] and (c) concert hall [41, 42]. Additionally, (d) the virtual reality (VR) scene with the characters and a participant during the experiment.

**Abstract**— In Immersive Audio Augmented Reality, a virtual sound source should be indistinguishable from the existing real ones. This property can be evaluated with the co-immersion criterion, which encompasses scenes constituted by arbitrary configurations of real and virtual objects. Thus, we introduce the term Audio Augmented Virtuality (AAV) to describe a fully virtual environment consisting of auditory content captured from the real world, augmented by synthetic sound generation. We propose an experimental design in AAV investigating how simplified late reverberation (LR) affects the co-immersion of a sound source. Participants listened to simultaneous virtual speakers dynamically rendered through spatial Room Impulse Responses, and were asked to detect the presence of an impostor, i.e., a speaker rendered with one of two simplified LR conditions. Detection rates were found to be close to chance level, especially for one condition, suggesting a limited influence on co-immersion of the simplified LR in the evaluated AAV scenes. This methodology can be straightforwardly extended and applied to different acoustics scenes, complexities, i.e., the number of simultaneous speakers, and rendering parameters in order to further investigate the requirements for immersive audio technologies in AAR and AAV applications.

**Index Terms**—Audio Augmented Virtuality, Co-immersion, Dynamic Binaural Synthesis, Reverberation, Virtual Acoustics

## 1 INTRODUCTION

The Reality-Virtuality Continuum [49] can be defined as a segment that spans between real and virtual environments. Augmented Reality (AR) and Augmented Virtuality (AV) can be situated along this continuum. AR is closer to the real world while AV is closer to a pure virtual environment, and results from capturing real-world content and bringing it into virtual reality (VR). Jerald [36, Ch. 21] provides examples of such real-world content, including 360° images, true 3D data captured via depth cameras or scanners, scientific data (e.g., volumetric datasets), and so on. We extend the AV definition to the acoustic domain and propose the term *Audio Augmented Virtuality* (AAV) as the creation of virtual environments using real-world auditory content. This is the first contribution explicitly working on AAV to the best of our knowledge.

Specifically, we propose an innovative experimental design in AAV aimed at evaluating how different acoustic reverberation approaches affect the *co-immersion* of concurrent speakers rendered in the same virtual environment. In the context of Audio Augmented Reality (AAR, [70]), co-immersion is defined as the property of simulated sound sources to be perceived as belonging to a real auditory scene,

rather than artificially superimposed upon it [59]. One possible approach to quantify co-immersion in an experimental setting is to measure users' ability to discriminate a virtual target from the real scene [66]. Our aim is to investigate co-immersion in a fully virtual context. This allows for more robust and repeatable experimental settings with respect to an AAR scenario involving real sound sources, as the same AAV scene can be experienced in any real environment using only a pair of headphones. In this case, co-immersion can be operationalized by asking users to discriminate among synthetic targets treated with different rendering approaches.

Realistic simulation of a virtual sound source via headphones poses several challenges, which are addressed by the literature on *dynamic binaural auralization* [7, 28]. *Auralization*, is the process of rendering audible the soundfield of a source in an acoustic space, simulating the listening experience at a given position in the modeled space [39]: a simple auralization approach is the convolution of an anechoic audio signal, representing a sound source, with a Room Impulse Response (RIR) recorded in the environment of interest. *Binaural* refers to the generation of two sound signals representing the sound waves reaching the listener's left-right ear canals. This entails accounting for the effects of the listener's body (head, torso, pinnae) interacting with incoming sound wavefronts based on their direction of arrival: Head-Related Transfer Functions (HRTFs) model this interaction. *Dynamic* means that the virtual simulation is adapted as the listener moves in the environment. Head rotations are commonly tracked with 3 degrees of freedom (DoF), although 6-DoF simulations are also possible.

Spatial RIRs, e.g. in Ambisonics format [71], allow for dynamic simulations by encoding directional properties of the soundfield reaching the listening point. Therefore, proper use of HRTFs with spatial RIRs provides an interactive and immersive simulation where a sound

- Davide Fantini, Giorgio Presti, Riccardo Bona and Federico Avanzini are with the University of Milan. E-mail: {davide.fantini | giorgio.presti | riccardo.bona | federico.avanzini}@unimi.it
- Michele Geronazzo is with the University of Padua and Imperial College London. E-mail: michele.geronazzo@unipd.it
- Alessandro Giuseppe Privitera is with the University of Udine. E-mail: privitera.alessandrogiuseppe@spes.uniud.it

Manuscript received 25 March 2023; revised 17 June 2023; accepted 7 July 2023.  
Date of publication 2 October 2023; date of current version 31 October 2023.  
This article has supplementary downloadable material available at <https://doi.org/10.1109/TVCG.2023.3320213>, provided by the authors.  
Digital Object Identifier no. 10.1109/TVCG.2023.3320213

source is perceived as stable in the Virtual Acoustic Environment (VAE) while the listener’s head rotates. In this work, we employ spatial RIRs recorded in real acoustic spaces, which can be conceptually equated to the capture of a 360° image in visual AV. When an RIR is brought into the virtual domain, listeners can explore the auditory scene as if they were in the place where the response was recorded.

We designed an experiment in which participants listened to 2, 3, or 4 simultaneous virtual speakers, rendered with dynamic binaural auralization through individualized HRTFs and Higher-Order Ambisonics (HOA) RIRs of real acoustic spaces. Each trial included one or no “impostor”, i.e. a speaker rendered with a simplified RIR with respect to the reference response, and participants were asked to detect and identify the impostor. Two approaches for RIR simplification were tested. In both cases, the direct sound and the early reflections (ERs) were kept unaltered with respect to the recorded responses, while two conditions were considered for the late reverberation (LR) part: (i) a static binaural downmix from the HOA response, and (ii) an artificial reverberator whose parameters were automatically tuned to match condition (i) using a recently proposed method [14]. Thus, in both conditions the simplified LR versions were static, i.e. their rendering was unaffected by the head movements. The rationale is that preserving dynamic spatial information in the ERs maintains sound source localization [17] and spatial impression [46, 47], whereas directional information in the LR should have limited perceptual influence [21, 44].

We evaluated the co-immersion of impostor speakers rendered using the two simplified LR approaches through the proposed experimental design in AAV. We analyzed the simplified RIRs both objectively, through quantitative features accounting for the accuracy of their fit to the reference responses, and subjectively, by investigating the experiment results using Signal Detection Theory (SDT). The goal of the evaluation can be summarized into the following research questions:

**Q1** How does a static LR affect the co-immersion of a sound source in an AAV scene?

**Q2** How is the co-immersion in an AAV scene influenced if the static LR of **Q1** is approximated with an artificial reverberator?

Condition (i) and (ii) are designed to address **Q1** and **Q2**, respectively. The conducted experiment allows us to evaluate the previously proposed automatic late reverberation matching method [14] in a co-immersion scenario. To our knowledge, this is the first evaluation of synthetic LR in a mixed reality scenario, a literature gap previously pointed out [53]. We provide the materials to replicate the experiment (both VR scene and auditory stimuli) in a public repository [23].

The paper is organized as follows. **Sec. 2** provides an overview of related literature’s experiments and criteria to evaluate the perception of virtual sound sources. In **Sec. 3**, we explain how the simplified LR conditions are obtained. **Sec. 4** describes the experiment materials and the methods used to design and evaluate the experiment. In **Sec. 5**, we report the obtained results which are discussed in **Sec. 6** with the experiment’s limitations. **Sec. 7** provides a conclusion of the paper.

## 2 BACKGROUND

Several criteria have been proposed to evaluate whether a simulated sound source is perceived as realistic. Among them, *authenticity* [13] requires the perceptual identity of the simulation with a real external reference. A virtual sound source is authentic if indistinguishable from a corresponding real one in a direct comparison. Authenticity is the most strict criterion since even small differences can be noticed by listeners. Brinkmann et al. [16] investigated the authenticity of individual dynamic binaural simulations in an ABX test. Real stimuli were reproduced with loudspeakers, while virtual sources were simulated through headphones. Results varied based on stimulus—speech was less authentic than noise—and environment, but not source position.

Only a few AAR applications require the simulation’s authenticity, thus less demanding evaluation criteria exist. One of them is *plausibility*, i.e. the simulation’s agreement with the listener’s expectation. Plausibility relies on the listener’s internal representation of reality, without an external reference as a comparison. Lindau and Weinzierl [45] proposed an experimental procedure to assess plausibility in VAEs by means of a 2AFC test with “real” and “simulation”

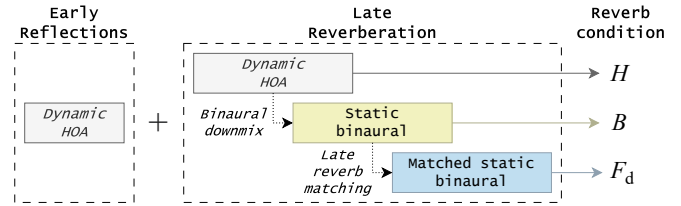


Fig. 2: Scheme of the reverberation conditions  $H$ ,  $B$  and  $F_d$ .

as choices. Participants listened to random trials of real stimuli and non-individual dynamic binaural simulations of them. Using SDT analysis, authors found a slight sensory difference between simulation and reality, though it was insufficient to represent a meaningful effect. A similar experiment was conducted by Pike et al. [57] to evaluate the plausibility of simulated stimuli with SDT. The authors found that participants’ sensitivity was not significantly greater than the minimum meaningful effect. Neidhardt et al. [52] investigated the influence of different BRIR simplification approaches on the plausibility in an interactive approaching motion towards a virtual loudspeaker rendered with dynamic binaural synthesis. Neidhardt and Zerlik [54] conducted a two-part plausibility experiment in 6 DoF. In the first part, where all stimuli were simulated, inexperienced listeners tend to accept virtual simulation as real, while experienced listeners’ responses were equally distributed. In the second part, similar to the experiment of Lindau and Weinzierl [45], results varied for inexperienced listeners, while experienced ones could reliably detect the simulation. Several more studies investigated the plausibility of simplified binaural rendering approaches [3]. Some of them employed VR environments [20] and analyzed the role of visual cues, too [6, 10]. Also, the authenticity and plausibility of different HRTF recording methods were assessed [55].

Between the strict authenticity based on direct comparison and the less demanding plausibility relying only on the listener’s expectation, there is another common scenario in AAR applications. In this scenario, real and virtual sound sources simultaneously reproduce different but similar stimuli. Thus, such a scenario allows a comparison of the virtual sound source with real ones but it prevents the direct comparison with the real counterpart. Two worth citing criteria for this scenario are *co-immersion* [66] and *transfer-plausibility* [69]. In a scene with existing sources, both these criteria evaluate the immersion of a simulated source without being perceived as such. Co-immersion and transfer-plausibility differ in their application. While transfer-plausibility is defined only for pure AAR scenes, co-immersion also encompasses fully artificial scenes where a sound source should be immersed among existing virtual sources by matching their acoustic features. Wirler et al. [69] evaluated the transfer-plausibility in a room with up to eight loudspeakers. None or one of them was simulated with dynamic non-individualized binaural synthesis. Results varied according to the number of simultaneous sources and stimulus type. Stecker et al. [66] investigated the co-immersion of a virtual speaker, represented by an avatar in a VR environment, by varying its acoustic features (room size and reflection coefficients) from the other speakers, virtual as well.

## 3 FROM AMBISONICS TO STATIC LATE REVERBERATION

In our study, we evaluated two approaches for simplifying the LR of HOA RIRs. In the remainder of the paper, the two simplified conditions will be referred to as  $B$ , and  $F_d$ , while the reference HOA one as  $H$ . **Fig. 2** provides a scheme of the three conditions: all use the reference ERs in HOA format—which can be dynamically rendered based on head rotations—whereas the LR part differs across conditions. For reference condition  $H$ , the full HOA RIR is used. For condition  $B$ , a simplified LR is obtained through a binaural downmix from HOA using the HRTFs of a Neumann KU100 dummy head [11] pointing towards the sound source. Therefore the resulting LR for  $B$  is a stereo response and is static with respect to head rotations. For condition  $F_d$ , the LR is generated algorithmically by the Freeverb artificial reverberator [63], a computationally efficient Schroeder reverberator. Freeverb’s parameters

are tuned by a LR matching method [14], using  $B$  as a target.

We made some changes to the matching method [14] in order to adapt it to the current context. In particular, we dismissed the ERs modeling and focused on LR exclusively. Then, we adapted the method to handle HOA instead of stereo RIRs. Moreover, in order to allow for a fair comparison, we applied a multichannel decorrelator based on randomized time-frequency delays and cascaded all-pass filters on the left and right channels of the artificial reverberator [40]. This is needed since stereo width in the original Freeverb algorithm is achieved by delaying one channel with a constant delay, thus introducing a perceivable pattern of correlated frequencies. The artificial reverberator resulting from our modification is a decorrelated Freeverb.

Fig. 3 depicts the block diagram to obtain the reverberation conditions. The input target HOA RIR  $R^h$  represents the reference condition  $H$ . As a first step, the binaural downmix procedure converts  $R^h$  into a corresponding binaural RIR  $R^b$  using the dummy head HRTF. Next, the boundary point  $\hat{\tau}$  between the ERs and LR of  $R^b$  is estimated, which is needed to extract the two RIR's parts using cosine fade-out and fade-in operations. Thus, we extracted the ERs  $E^h$  from the target HOA RIR  $R^h$  with a cosine fade-out 128 samples long ending at  $\hat{\tau}$ . Then, the LR  $L^b$  is obtained through a cosine fade-in on  $R^b$ . Condition  $B$  is represented by the combination of the HOA ERs  $E^h$  and the stereo LR  $L^b$ . In the LR matching step, we obtained the approximated LR  $\hat{L}$  by tuning Freeverb's parameters to match  $L^b$ . Finally, condition  $F_d$  is represented by the combination of the HOA ERs  $E^h$  and the LR  $\hat{L}$  followed by the decorrelation procedure. The ERs and LR separation and the LR matching methods are described in Sec. 3.1 and Sec. 3.2, respectively.

### 3.1 Early reflections and late reverberation separation

A custom method is used to estimate the boundary point  $\hat{\tau}$  between ERs and LR of the RIR  $R^b$ . Such signal is divided into multiple time frames where the  $n$ -th frame ranges from  $R^b[0]$  to  $R^b[n\gamma - 1]$ , with  $\gamma = 512$ . A cosine fade-out is applied at the end of each frame. Then, we fit an autoregressive model of order  $2\lfloor(2 + f_s/1000)\rfloor$  for each frame  $n$  using the Yule-Walker method. The obtained autoregressive coefficients  $A_n$  are used to compute the Power Spectral Density  $PSD_n$ . Next, the Logarithmic Spectral Distance (LSD) is computed for each pair of consecutive PSD values  $PSD_n$  and  $PSD_{n+1}$ . A knee detection algorithm [60] is applied to the LSD values to estimate the frame containing standalone ERs, which last sample is set as the boundary point  $\hat{\tau}$ . The rationale is that ERs frames yield higher differences of PSD values compared to those containing diffuse reverberation as well.

### 3.2 Artificial late reverberation matching

The core procedure for generating condition  $F_d$  is the LR matching procedure depicted in Fig. 3, where Freeverb's parameters are automatically tuned to minimize the difference between the target binaural LR  $L^b$  and the artificially generated one  $\hat{L}$ . To obtain  $\hat{L}$ , a Freeverb's impulse response is generated given a set of parameters  $P$ , then a cosine fade-in from 0 to  $\hat{\tau}$  is applied to remove the ERs. Bayesian optimization is employed to tune the parameters using a Gaussian process as a prior [64]. The objective is to minimize the loss function  $\ell$  between the signals  $L_s^b$  and  $\hat{L}_s$ , obtained by convolving with a 1 s long logarithmic sweep the signals  $L^b$  and  $\hat{L}$ , respectively. Computing  $\ell$  on sweeps instead of impulses provides a better match of the LR gain. Loss functions used for similar tasks [22, 43] inspired the definition of  $\ell$  as the mean absolute difference between the multi-resolution mel-spectrograms of  $L_s^b$  and  $\hat{L}_s$ . The mel-spectrogram in dB  $\mathcal{M}_f$ —Short-Time Fourier Transform, followed by the filter-bank mapping to the mel scale and dB conversion—aims at a perceptually motivated loss. For the STFT, we used a Hanning window of size  $f$  with a 25% overlap. The loss function  $\ell$  is so defined:

$$\ell(L_s^b, \hat{L}_s) = \sum_{f \in F} \frac{1}{T} \sum_{t=1}^T \left| \mathcal{M}_f^t(L_s^b) - \mathcal{M}_f^t(\hat{L}_s) \right|, \quad (1)$$

where  $t$  is the time frame index,  $T$  is the number of frames and  $F = \{256, 512, 1024, 2048, 4096\}$  are the selected spectrum frame sizes.

Before the computation of  $\ell$ , we cut infrasonic components from  $L_s^b$  and  $\hat{L}_s$  with a third-order high-pass filter. Further, the spectrum's values are truncated to the lower threshold of -60 dB. We computed the average of  $\ell$  values for the left and right channels to obtain a unique value.

The loss function  $\ell$  is assumed to follow a multivariate Gaussian distribution. An acquisition function  $\mathcal{A}$  selects the next set of parameters  $P_{i+1}$  from a range of possible values. The prior Gaussian distribution of  $\ell$  is modeled employing the Matérn kernel [68, Ch. 4, Sec. 4.2] as a covariance function between the current parameters  $P_i$  and the new candidate ones. Thus, Freeverb's parameters are iteratively tuned to minimize  $\ell$  by means of the Bayesian optimization.

Finally, condition  $F_d$  is obtained by recombining the original HOA ERs  $E^h$  and the matched stereo LR  $\hat{L}$ , followed by the above described decorrelation.

## 4 MATERIALS AND METHODS

### 4.1 Virtual Acoustic Environments

The three VAEs evaluated in our experiment consist of datasets of HOA RIRs recorded in a living room [61], a classroom [56] and a concert hall [41, 42]. These datasets contain recordings of at least four sound sources in fixed positions. Each RIR was recorded at a 48 kHz sample rate using an em32 Eigenmike. We specifically selected these rooms as they represent the large variability in room acoustic properties.

The **living room** [61] is a small environment of  $4.97 \times 3.78 \times 2.71$  meters (width  $\times$  length  $\times$  height) that includes furniture such as a sofa, armchairs, a carpet, etc. As a result, it has a short reverberation time. From this dataset, we selected the RIRs recorded in position R1. The selected sound sources were S5, STV, SG4 and SG6 (see Fig. 1a), which we renamed as  $LI_{-45^\circ}$ ,  $LI_{0^\circ}$ ,  $LI_{+39^\circ}$  and  $LI_{+66^\circ}$ , respectively, to clarify the angle between the receiver and the source. These sources were placed between 1.6 m and 2.6 m from the receiver.

The **classroom** [56] is an empty room, slightly larger than the living room ( $6.5 \times 8.3 \times 2.9$  m), resulting in a longer reverberation time. The recordings were made with the receiver in the middle of a 3D grid representing the recording positions of the sound sources. We selected the four RIRs recorded at the receiver's height, i.e. 1.5 m, for the source positions labeled 152, 032, 112, and 302 in the dataset (see Fig. 1b). In this paper, they are referred to as  $CL_{-45^\circ}$ ,  $CL_{0^\circ}$ ,  $CL_{+45^\circ}$ , and  $CL_{+90^\circ}$ , respectively. The source distances from the receiver were about 1.4 m for  $CL_{-45^\circ}$  and  $CL_{+45^\circ}$  and 1.5 m for  $CL_{0^\circ}$  and  $CL_{+90^\circ}$ .

The **concert hall** [42], originally a church, is the largest environment among the three, resulting in a long reverberation time. The RIRs were recorded with the loudspeakers on the stage arranged as an ensemble and the receiver placed in front of them. We selected the four positions of the sound source recorded 3 m away and at the angles  $-30^\circ$ ,  $0^\circ$ ,  $+30^\circ$ , and  $+90^\circ$  with respect to the receiver (see Fig. 1c). We named these positions  $CO_{-30^\circ}$ ,  $CO_{0^\circ}$ ,  $CO_{+30^\circ}$ , and  $CO_{+90^\circ}$ , respectively.

The VAEs were rendered in the same VR scene designed in Unity. The VR scene is a mostly dark environment with a spotlight on the floor, as shown in Fig. 1d. We designed the scene to be as simple as possible to minimize possible biases elicited by the visible environment, and at the same time to provide the participants with coherent sensory-motor contingencies for head rotations.

We computed a set of acoustic parameters to provide an objective acoustic characterization of the VAEs aimed at identifying the differences between VAEs and the various positions of the sound source within each VAEs. We computed these parameters considering the binaural downmixed versions of the HOA RIRs  $R^h$  with the listener's head oriented toward the frontal position. For this downmix, we used again the HRTF of the Neumann KU100 dummy head [11].

The selected acoustic parameters are defined in ISO standard 3382 [26] and include reverberation time  $T_{20}$ , early decay time  $EDT$ , center time  $T_5$ , clarity index  $C_{80}$ , lateral energy fraction  $LF_{80}$ , and interaural cross-correlation coefficient  $IACC$  [32]. Reverberation time is the time it takes for the reverberation to decay by 60 dB after the sound source stops. In particular, we measured the reverberation time  $T_{20}$  by tripling the milliseconds employed by the RIR to decay from  $-5$  to  $-25$  dB. Similarly, early decay time  $EDT$  measures the millise-





Table 1: Acoustic characterization through a set of parameters (ISO 3382 [26]) computed with AURORA [24, 25] for each position in the VAEs.

VAE	Position	$T_{20}$ [ms]	$EDT$ [ms]	$T_S$ [ms]	$C_{80}$ [dB]	$LF_{80}$ [dB]	$IAAC$
Living Room	$LI_0^\circ$	404.5	401.0	19.6	13.9	-6.4	0.599
	$LI_{+39^\circ}$	425.5	367.0	15.3	14.5	-1.4	0.267
	$LI_{-45^\circ}$	390.5	402.5	19.3	13.7	-0.5	0.229
	$LI_{+66^\circ}$	396.0	375.0	14.4	15.9	0.2	0.143
Classroom	$CL_0^\circ$	1057.1	828.5	38.1	8.2	-1.2	0.732
	$CL_{-45^\circ}$	1178.1	951.5	45.1	7.3	-1.4	0.240
	$CL_{+45^\circ}$	1152.4	1003.0	44.3	7.5	-0.8	0.223
	$CL_{+90^\circ}$	1061.4	950.0	41.4	7.8	-1.0	0.224
Concert Hall	$CO_0^\circ$	1635.5	19.5	6.6	15.8	-10.3	0.809
	$CO_{-30^\circ}$	1654.0	479.5	14.2	12.9	-1.2	0.379
	$CO_{+30^\circ}$	1640.0	461.5	12.4	13.6	0.0	0.401
	$CO_{+90^\circ}$	1594.5	432.5	23.4	12.5	-0.7	0.157

where  $c$  is the speed of sound, the head radius  $a$  is computed as  $a = (0.41 \frac{X1}{2} + 0.22 \frac{X3}{2} + 3.7)$ , and  $\gamma$  is the angle between the source vector  $\vec{S}$  and the ear vector  $\vec{e}$  with origin in the center of the sphere.

To model the high-frequency content, the best-fit HRTF is selected from the CIPIC [2] dataset according to anthropometric features. For each participant, we manually annotated  $n = 15$  ear contours  $C1$  and  $k = 20$  ear canal positions. These annotations are used to estimate the pinna notch frequencies  $f_0$  for each elevation angle  $\phi \in \{1, \dots, N_\phi\}$  [27]. The HRTF selection is based on the mismatch between the notch frequencies  $f_0$  and the ones  $F_0$  extracted from the CIPIC HRTFs:

$$m_{(k,n)} = \frac{1}{N_\phi} \sum_{\phi} \frac{|f_0^{(k,n)}(\phi) - F_0(\phi)|}{F_0(\phi)}. \quad (3)$$

### 4.3 Objective analysis

We objectively evaluated the LR matching method (see Sec. 3) by comparing the acoustical parameters of the target reverberation and the matched one generated by Freeverb with the automatically tuned parameters. The target RIR is  $R^b$ , i.e. the binaural downmix of the HOA RIR  $R^h$  with the listener's head oriented toward the sound source. For the matched reverberation, the acoustical parameters cannot be computed on matched LR  $\hat{L}$  alone. Therefore, to obtain the complete matched RIR  $\hat{R}^b$ , we joined  $\hat{L}$  with the ERs of  $R^b$  using a cross-fade centered on  $\hat{\tau}$ . For this objective evaluation, we selected the reverberation time  $T_{20}$ , the center time  $T_S$ , and the clarity index  $C_{80}$  as acoustical parameters. We excluded  $EDT$  and  $LF_{80}$  since they are specifically defined for ERs, which are the same for the target and the matched RIRs compared in this evaluation. Additionally, we also included the loss function value  $\ell_{end}$  at the end of the iterative LR matching procedure.

### 4.4 Subjective Evaluation

We evaluated the co-immersion of conditions  $B$  and  $F_d$  compared to the reference reverberation  $H$  conducting a listening experiment with human participants in AAV. For each participant, we first took the pictures needed for the HRTF individualization procedure. Then, we explained the experiment procedure and we had the participant wear the VR headset and headphones. The procedure involved a preliminary screening test followed by the main experiment. During the experiment, each participant performed two tasks while experiencing the dynamic auralization of the VAEs where up to four simultaneous speakers were rendered. Participants were seated in a soundproof room (weighted sound reduction index  $R_w \geq 68$  dB) and were able and encouraged to move their heads and torso during each trial.

#### 4.4.1 Stimuli

The speech data used in our experiment were obtained from the ACE challenge's corpus [19], which includes English speeches recorded in an anechoic chamber at a sample rate of 48 kHz and 16-bit depth.

Speech signals were normalized to a target loudness of  $-23$  LUFS. The four speakers used in our experiment were two males and two females—with a mix of native and non-native English speakers—retrieved from the ACE corpus. Each of their speeches was about 1 minute long.

The four speakers were associated with four characters from the *Guess Who?* board game. Additionally, for each character, we created a short backstory that was coherent with their utterances. In the preliminary screening test, we introduced each character with their name and background. During the preliminary test, in the VR scene, we showed the characters' faces in the voices' direction while the participant listened to them (see Fig. 1d). Using backstories and faces, we aimed to provide a more complete characterization of the speakers. This was intended to help participants remember the name of each speaker, which is needed in the main experiment where the characters' faces were hidden. However, we allowed participants to identify a speaker using other features, such as physical traits and backstory insights.

#### 4.4.2 Experiment design

We designed our listening experiment to evaluate the co-immersion of the simplified LR approaches in an AAV scene. In particular, we compared the perception of the three reverberation conditions described in Sec. 3:  $H$ ,  $B$  and  $F_d$ . Each trial of the experiment involved  $q$  speakers, where  $q \in \{2, 3, 4\}$ . The speakers were placed in different positions of one of the three VAEs (living room, classroom, concert hall). In each trial, we provided the participants with one of three scenarios None, Bin and Fv, related to the reverberation conditions  $H$ ,  $B$  and  $F_d$ , respectively. In the None scenario, all  $q$  speakers are rendered with the  $H$  reverb condition. In scenarios Bin and Fv,  $q - 1$  speakers are rendered with  $H$  condition, while the remaining speaker is rendered with  $B$  and  $F_d$  conditions, respectively. Speakers rendered with  $B$  and  $F_d$  conditions are called *impostors*. Following SDT terminology, we refer to the presence of an impostor as *signal* and the absence as *noise*.

For each trial, participants perform the following two tasks to assess the co-immersion of the impostor conditions:

1. *Detection*: participants were asked to detect the presence (signal) or absence (noise) of an impostor. Thus, we asked the following question: "Are all the speakers talking in the same room?"
2. *Identification*: only if the participants answered affirmatively to the previous question, they were asked to identify which of the  $q$  speakers was the impostor. This task was represented by the question: "Which of the speakers do you think is not in the same room as the others?". Participants answered this question with the name of the impostor. When  $q = 2$ , we decided to disregard the identification task since the impostor's identification between two speakers is not possible without other references.

To guide participants in correctly interpreting the tasks, we provided some recommendations before the start of the experiment and we repeated them upon request. Participants should focus on the acoustic and reverberation of the speeches rather than the spatial position and distance of the speakers. We recommended detecting the presence

Table 2: Objective evaluation through acoustical parameters of the reverb matching method averaged across the positions of each VAE. For each metric ( $\ell_{end}$  excluded), we report in column  $R^b$  the mean and the standard deviation of the metrics' value for the target RIRs  $R^b$ . In column  $\hat{R}^b - R^b$ , we report the mean and the standard deviation of the signed error between the metric values for the matched RIRs  $\hat{R}^b$  and the target RIRs  $R^b$ . In this column, we also report between brackets the percentage of the mean error with respect to the mean metric value of  $R^b$ .

VAE	$\ell_{end}$ [dB]	$T_{20}$ [ms]		$T_S$ [ms]		$C_{80}$ [dB]	
		$R^b$	$\hat{R}^b - R^b$ (%)	$R^b$	$\hat{R}^b - R^b$ (%)	$R^b$	$\hat{R}^b - R^b$ (%)
Living Room	$3.5 \pm 0.2$	$405.0 \pm 4.7$	$-110.9 \pm 20.0$ (-27.4)	$18.8 \pm 3.5$	$-1.2 \pm 0.6$ (-6.5)	$14.2 \pm 0.9$	$3.3 \pm 0.9$ (23.3)
Classroom	$8.7 \pm 0.1$	$1163.7 \pm 31.5$	$-196.8 \pm 40.5$ (-16.9)	$40.1 \pm 3.5$	$-14.4 \pm 1.5$ (-35.9)	$7.6 \pm 0.5$	$3.1 \pm 0.3$ (40.9)
Concert Hall	$6.7 \pm 0.5$	$1629.0 \pm 11.8$	$-256.2 \pm 62.3$ (-15.7)	$10.8 \pm 2.7$	$-0.2 \pm 0.2$ (-1.6)	$13.8 \pm 1.2$	$-0.4 \pm 0.1$ (-2.9)

of an impostor only if they thought that the acoustic rendering of a speaker was incoherent with the others. We also suggested exploring the acoustic scene in all directions by moving their head. We told the participants that the impostor's occurrence across the trials was random—though their occurrence followed a precise protocol—and we informed them in which of the three VAEs the simulation was.

#### 4.4.3 Preliminary screening test

To ensure the ecological validity and reliability of the participants' judgments experiencing the proposed VAEs, we conducted a preliminary screening test on sound externalization. The sense of presence refers to the feeling of being physically in the virtual environment [62]. Sound externalization is an important aspect of it because refers to the ability to accurately create the sensation that the sound is coming from specific external locations, rather than just being heard through headphones [12]. In each trial of this test, participants listened to the speech of a single speaker rendered in a given VAE with the character's face visible. Then, they were asked to rate the degree of perceived externalization between three options: inside, on the edge, or outside the head. Each speaker was repeated for three trials, one for each of the three reverberation conditions  $H$ ,  $B$ , and  $F_d$ . We excluded from the experiment the participants who rated condition  $H$  as inside the head for at least 50% of its occurrences. In addition to screening purposes, the preliminary test also served to acquaint participants with the VAEs and the VR scene, as well as to learn the association between the speakers' voices and their names, which was necessary for the main experiment. Thus, after the preliminary test, we asked them to guess the speaker's name for each speech without showing the character's face.

#### 4.4.4 Protocol

The experiment consisted of 27 trials, obtained by combining three reverberation scenarios (None, Bin and Fv), three VAEs (living room, classroom and concert hall), and three complexities (2, 3 and 4). The trials were divided into three sessions, one for each complexity in Latin square order. Participants were asked to take a break between each session. Within each session, we randomly assigned the VAEs order and, for each of them, we presented three consecutive trials, one for each scenario, in random order.

For each complexity, we selected a set of predetermined source positions in each VAE to render the speakers. The position of the impostor was also predetermined but varied based on the VAE, complexity, and the Bin and Fv scenarios. The complete list of the selected positions is reported in the supplementary materials. For each trial, the involved speakers were randomly selected and assigned to the predetermined source positions. Thus, even though the impostor's position was predetermined, the speaker acting as the impostor was randomly selected.

In the preliminary screening test, each speaker was consecutively repeated for each reverberation condition. With the four selected speakers, the preliminary test resulted in 12 trials. Each speaker was placed in a source position of a VAE that was the same for all participants. The speakers were presented in a Latin square order, while the reverberation conditions were presented randomly within each speaker's trials.

#### 4.4.5 Participants

We recruited 31 participants to take part in our experiment. One participant reported experiencing tinnitus and was therefore excluded from

the analysis. Other 5 participants failed the preliminary screening test; therefore, they were excluded. The remaining 25 participants (17 males, 7 females, 1 not specified) reported normal hearing conditions and passed the screening test, thus they were included in the analysis. The age distribution of the participants was: 17 in 18–27, 7 in 28–37, and 1 in 38–47. Most of the participants (16) reported an intermediate level of experience with audio reverberation, 7 of them had no experience and the remaining 2 were self-reported experts. Regarding VR, most of the participants (21) had no experience, 3 reported an intermediate level and the remaining 1 expert level. Participants completed the experiment, including the preliminary test, between 30 and 80 minutes.

## 5 RESULTS

### 5.1 Matching accuracy

In this section, we discuss the results of the objective analysis presented in Sec. 4.3, which are reported in Tab. 2 averaged across the positions of each VAE. The complete values for all positions are reported in the supplementary materials. Since  $\ell_{end}$  exhibited small standard deviation values, positions inside each VAE are similarly matched. On the other hand, we noticed differences in the  $\ell_{end}$  values of the VAEs, characterizing a dependency on the type of RIR used. In particular, shorter RIRs, such as the ones in the living room, tend to yield lower values. For the sake of comparability, acoustical features can provide a more reliable evaluation. Reverberation time  $T_{20}$  is underestimated in each VAE, as well as the center time  $T_S$ , especially in the classroom. Moreover, the classroom exhibits the worst performance for clarity  $C_{80}$ , with a mean percentage error of 40.9%. Nonetheless, the error standard deviations are in general quite low, suggesting that the differences in the matching accuracy between the positions inside each VAE are limited.

### 5.2 Preliminary test

Fig. 5 displays the frequencies of the externalization rates collected in the preliminary test grouped by reverberation condition. Stimuli are more frequently perceived outside the head compared to the other options. As expected, this tendency is emphasized for the reference reverberation condition  $H$  compared to the impostor's conditions  $B$  and  $F_d$ . The condition  $B$  is more frequently perceived inside the head than  $F_d$ , which could be a consequence of the decorrelation operation applied to Freeverb's output. Despite that, we also notice that  $B$  is perceived as outside the head more frequently than  $F_d$ . The latter has received more "on the edge" responses than the other conditions. However, the general tendency toward externalization satisfies our requirement of providing participants with ecologically spatialized stimuli.

### 5.3 Experiment results analysis

Fig. 6 shows the alluvial plot summarizing the overall participants' performances in the detection and identification tasks. As expected, Fv scenario is more detectable than Bin based on their hit rates—percentage of correct signal responses—of 64.4% and 54.7%, respectively (see streams from the Fv and Bin nodes to the Signal one in Fig. 6). Fv is also more identifiable according to the percentage of correct identification responses among the Fv (33.8%) and Bin trials (22.2%). Despite the higher hit rate for Fv, Pearson's  $\chi^2$  test showed no statistically significant difference between the three scenarios ( $\chi^2 = 4.50$ ,  $p = 0.11$ ,  $df = 2$ ) in predicting if the detection response was wrong or correct.



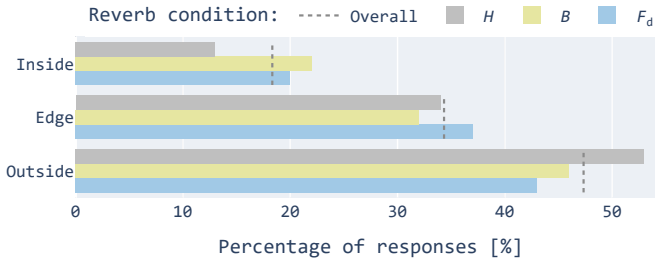


Fig. 5: Externalization rates frequencies in the preliminary screening test grouped by reverberation condition.

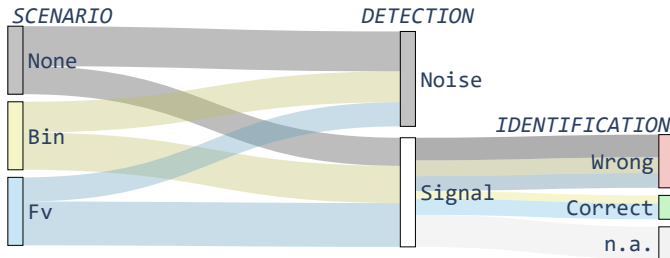


Fig. 6: Alluvial plot of the percentages of responses in the detection and identification tasks. The disregarded identification responses for complexity 2 are accounted in “n.a.”.

In addition to the overall analysis, we investigated the influence of complexity and VAE. The alluvial plots for each combination of complexity and VAE are reported in the supplementary materials. Tab. 3 reports the percentages of correct detection for each combination of scenario, complexity and VAE. We conducted a logistic regression test to assess the significance of the independent variables (scenario, complexity and VAE) influence and their interactions. The dependent binary variable was again the correct (1) or wrong (0) detection. We considered the combinations of up to two variables to limit the number of the resulting combinations. In Tab. 3, we report the significance level of the regression coefficients (log-odds) for each variables’ combination. A positive coefficient (light gray cells in Tab. 3) implies an increment in the predicted log-odds of the dependent variable and vice versa for a negative coefficient (dark gray cells in Tab. 3). Regarding all VAEs, we noticed that the correct rejection rate—the rate of correct noise responses for the None scenario—decreases as the complexity increases, implying an increment of the false alarm rate. This effect is supported by logistic regression since for the None scenario, the coefficient is significantly positive ( $p < .001$ ) and negative ( $p < .01$ ) for complexity 2 and 4, respectively. Fv scenario shows a reversed trend, i.e. the hit rate increases with the complexity. This is supported again by the negative and positive coefficients obtained in Fv for complexity 2 ( $p < .01$ ) and 4 ( $p < .01$ ), respectively. Thus, when an impostor is rendered with reverb condition  $F_d$ , the higher the complexity, the easier the detection task. Conversely, for Bin scenario the higher hit rate (60%) is found in complexity 2 and it drops for greater complexities.

Analyzing the role of VAEs, the classroom exhibits a significant positive coefficient ( $p < .001$ ) with an overall correct detection rate of 67.1%. In this VAE, the hit rate of Fv (80%) is associated with a significant positive coefficient ( $p < .01$ ). In contrast to the classroom, in the living room, Fv has a low hit rate (48%) with a significant negative coefficient ( $p < .01$ ). For Bin a significant negative coefficient ( $p < .05$ ) is found in the concert hall where the hit rate is 44.0%. These results suggest that the impostor is more detectable for Fv than Bin in the classroom and concert hall. The opposite pattern is noticed in the living room. However, in the overall case of Fv, a significant positive coefficient was found ( $p < .01$ ). Finally, another significant positive coefficient was found for the overall case of complexity 2 ( $p < .01$ ).

### 5.3.1 Identification task

For the identification task, the overall percentages reported in Sec. 5.3 are difficult to interpret since multiple complexities are considered. Thus, we computed the percentages of correct responses for each complexity, except for complexity 2 which was disregarded for the identification task. For complexity 3, 13.3% of the identification responses were correct among the Bin trials, against the 42.7% among the Fv trials. This difference decreases in complexity 4 where the percentages are 21.3% and 28% for Bin and Fv, respectively. It is worth noticing that for Bin scenario the percentage of correct identification is lower for complexity 3 than 4, although the inferior number of simultaneous speakers which entails a higher chance level. Furthermore, these percentages are even lower than the chance levels of both complexity 3 (33%) and 4 (25%). However, we avoid here any speculation to explain the trends obtained in the identification task. The proposed experimental design cannot reliably handle such complexity and it will be subject to revisions in future versions of the experiment.

### 5.3.2 Signal Detection Theory analysis

Signal Detection Theory (SDT) is a psychophysical approach to model performance in decision processes by measuring the ability to detect a certain information (signal) among noise [33]. We employed SDT to analyze the detection task results which falls under the yes/no paradigm. In our case, as already mentioned, the None scenario is regarded as *noise*, while Bin and Fv are regarded as *signals*. In SDT, the noise and signal conditions are modeled as Gaussian probability density distributions of equal variance depending on the subject’s internal response to the provided stimulus. The sensitivity index  $d'$ , defined as the distance of the two distributions, measures the ability to discriminate between the two stimuli. A  $d'$  close to 0 denotes the subject’s inability to discriminate between noise and signal conditions, thus, in our case, the impostor cannot be detected. Instead, a large  $d'$  denotes a high discrimination ability, in our case, the impostor can be easily detected. In SDT, the sensitivity  $d'$  is estimated as follows:

$$d' = z(p_{Hit}) - z(p_{FA}), \quad (4)$$

where  $z$  is the inverse cumulative normal distribution, while  $p_{Hit}$  is the hit rate and  $p_{FA}$  is the false alarm rate. Another metric provided by SDT is the criterion  $c$  which measures the response bias, i.e. the subject’s tendency toward one of the two responses. In our case, negative and positive values of  $c$  imply a tendency to detect the presence and the absence of an impostor, respectively. In SDT,  $c$  is estimated as follows:

Table 3: Percentages of correct detection for each scenario, complexity and VAE for the experiment’s participants. We also show the significant results of the logistic regression test providing the coefficient’s sign of the combination of variables (light gray for positive and dark gray for negative signs) and the significance level as superscript (\*:  $p < 0.05$ , \*\*:  $p < 0.01$ , \*\*\*:  $p < 0.001$ ).

		Complexity 2	Complexity 3	Complexity 4	All complexities
Living room	None	84.0	72.0	88.0	61.3
	Bin	68.0	48.0	48.0	54.7
	Fv	36.0	52.0	56.0	48.0**
	All	62.7	46.7	54.7	54.7
Classroom	None	68.0	56.0	44.0	56.0*
	Bin	68.0	64.0	64.0	65.3
	Fv	80.0	72.0	88.0	80.0**
	All	72.0	64.0	65.3	67.1***
Concert hall	None	76.0	64.0	36.0	58.7
	Bin	44.0	40.0	48.0	44.0*
	Fv	56.0	68.0	72.0	65.3
	All	58.7	57.3	52.0	56.0
All VAEs	None	76.0***	53.3	46.7**	58.7
	Bin	60.0	50.7	53.3	54.7
	Fv	57.3**	64.0	72.0**	64.4**
	All	64.4**	56.0	57.3	59.3

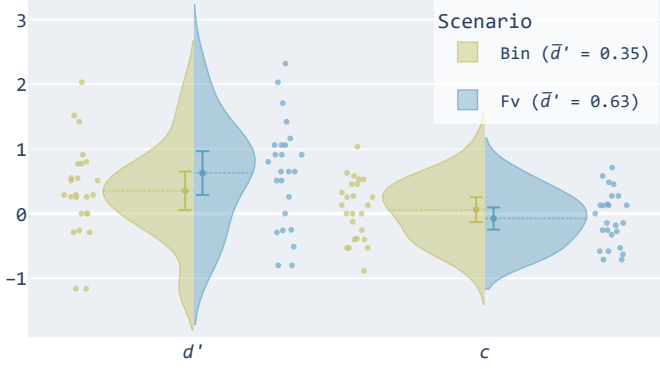


Fig. 7: Distributions of the SDT metrics  $d'$  and  $c$  for the Bin and Fv scenarios. The dashed horizontal lines represent the distribution's mean, while the vertical whiskers are the 95% confidence intervals.

$$c = -\frac{1}{2} (z(p_{Hit}) + z(p_{FA})). \quad (5)$$

To avoid the occurrence of infinite values in SDT metrics, we applied the log-linear correction that consists of adding 0.5 at both the numerator and denominator used for the computation of  $p_{Hit}$  and  $p_{FA}$  [65]. We computed the SDT metrics  $d'$  and  $c$  independently for each experiment's participant. Further, we separately computed the SDT metrics  $d'_{Bin}$  and  $c_{Bin}$  for None + Bin trials and the metrics  $d'_{Fv}$  and  $c_{Fv}$  for None + Fv trials. The distributions of such metrics are shown in Fig. 7. The mean values of  $d'$  distributions are  $\bar{d}'_{Bin} = 0.35$  (95% CI:[0.05, 0.65]) and  $\bar{d}'_{Fv} = 0.63$  (95% CI:[0.29, 0.97]). A sensitivity value  $d' = 0$  would correspond to the complete inability to detect the impostor. However, since in inferential statistics is impossible to directly prove the null hypothesis  $H_0$  that  $d' = 0$ , we rely on the minimum effect hypothesis [45, 51]. Thus, we try to reject the alternative hypothesis  $H_1$  that  $d' > d'_{min}$ , where  $d'_{min}$  is a lower threshold to regard the sensitivity as too small to be perceptually relevant. In previous works employing SDT to assess plausibility [6, 45, 57], the acoustic simulation would be considered plausible if the obtained detection rate was less than 55%, corresponding to a  $d'_{min} = 0.1777$ . Given this threshold and the selected type I and II error levels, the authors estimated the lower bound of the optimal sample size, under the assumption of unbiased subjects and an equal number of noise and signal trials. Since we recruited 25 participants in our experiment, each performing 27 trials, we collected 675 samples. However, our analysis is independent for the None + Bin and None + Fv trials, thus they account for  $N = 450$  samples each. Selecting the type I error level  $z_\alpha = .25$  and the type II error level  $z_\beta = .05$ , we estimated a  $d'_{min} = 0.274$  with the following equation:

$$d'_{min} = \sqrt{(z_\alpha + z_\beta)^2 \frac{2\pi}{N}} \quad (6)$$

To find if the sensitivity mean values  $\bar{d}'_{Bin}$  and  $\bar{d}'_{Fv}$  represent a meaningful effect, we first ascertained the normality of both  $d'_{Bin}$  and  $d'_{Fv}$  distributions through a Shapiro-Wilk test ( $p > .05$ ). Then, with a one-sided paired t-test, we found that  $\bar{d}'_{Fv}$  is significantly greater than  $d'_{min}$  ( $p < .05$ ), while for  $\bar{d}'_{Bin}$  there is no significance ( $p = 0.3$ ). Thus, the sensitivity for Bin is not sufficient to represent a meaningful effect. We also computed the mean values of the criterion metric  $\bar{c}_{Bin} = 0.06$  and  $\bar{c}_{Fv} = -0.08$ . We found that they are both not significantly different from 0 according to a t-test ( $p > .05$ )—both  $c_{Bin}$  and  $c_{Fv}$  were normally distributed according to a Shapiro-Wilk test. This result suggests that there is no clear tendency toward noise or signal response.

Further, we investigated the participants' performances independently for each complexity. The average  $d'$  values for each complexity are reported in Fig. 8a. We noticed that  $\bar{d}'_{Bin}$  decreases with the complexity, while  $\bar{d}'_{Fv}$  decreases from complexity 2 and complexity 3, but it is stable between complexities 3 and 4. Further,  $\bar{d}'_{Fv}$  is higher than  $\bar{d}'_{Bin}$

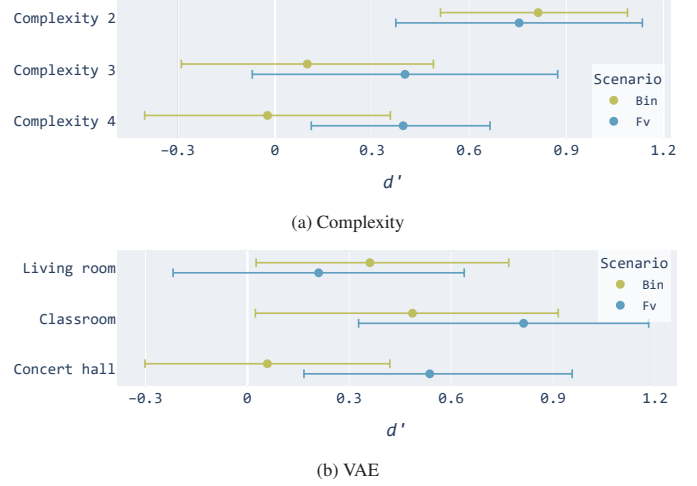


Fig. 8: Mean values  $\bar{d}'$  with 95% confidence intervals of Bin and Fv scenarios for each (a) complexity and (b) VAE.

for all complexities, except for complexity 2 where it is slightly lower. The statistical significance of these differences was assessed with a two-way ANOVA test regarding scenario (Bin and Fv) and complexity as factors. We found a significant effect on  $\bar{d}'$  only for complexity ( $F(2) = 6.47, p < .01$ ). Then, we conducted a Tukey-HSD post hoc test on the complexity levels showing a significant difference between complexities 2 and 3 ( $p < .05$ ) and complexities 2 and 4 ( $p < .01$ ).

We performed a similar analysis for the VAE influence on  $\bar{d}'$ . The average  $\bar{d}'$  values for each VAE along with their 95% confidence intervals are reported in Fig. 8b. The classroom is the VAE where the impostor's detection is easier, i.e. with the higher mean sensitivity, confirming the results reported in Tab. 3. Conversely, the Bin scenario is almost undetectable in the concert hall. We noticed that  $\bar{d}'_{Bin}$  is lower than  $\bar{d}'_{Fv}$ , except for the living room. However, no significant effect was found with a two-way ANOVA test with scenario and VAE as factors.

## 6 GENERAL DISCUSSION

In the literature, the Level Of Audio Detail (LOAD) [4] refers to the adaptation of the audio rendering complexity. LOAD techniques aim to reduce the rendered details to save computational cost while minimizing the perceived quality degradation. Accordingly, our study can be interpreted as an evaluation of the LOAD required for LR to preserve the perceived co-immersion in AAV scenes. The obtained promising results on simplified LR have a potentially relevant impact on the saving of computational and storage resources compared to convolution with RIR or wave and geometric acoustics-based approaches.

We found that the detection tasks for Bin and Fv scenarios are very close to the chance level both accounting for detection rates and SDT metrics. Such results suggest the participants' inability to reliably detect the impostor in the proposed experiment. For Bin scenario, we found that the mean sensitivity  $\bar{d}'_{Bin}$  is not significantly greater than  $d'_{min}$ , thus it does not represent a meaningful effect. This result suggests that the LOAD can be reduced by substituting the dynamic rendering of an HOA LR with a static binaural version. Thus, we can answer to Q1 that a static LR does not affect the co-immersion of a sound source in the evaluated AAV scenes. For Fv scenario, the mean sensitivity  $\bar{d}'_{Fv}$  is significantly greater than  $d'_{min}$  but is still very close to this value. To answer Q2, this finding suggests that the static LR can be approximated with a simple artificial reverberator, such as Freeverb, with a perceivable but limited influence on co-immersion. It is worthwhile to notice that the LOAD is further reduced in Fv scenario because artificial reverberators are more efficient compared to convolution. However, it comes at the cost of a greater impact than Bin on the perceived degradation. Finally, the LR matching method leads to satisfactory results if evaluated with the co-immersion criterion compared to the authenticity one reported by its authors [14]. Since in



their study the matched reverberation was clearly distinguishable from the reference in direct comparison, co-immersion can be considered a narrow criterion that can be exploited in specific applications.

Although the LR simplified versions are difficult to detect, we found important differences with respect to the overall case by analyzing separately each VAE and complexity. A significant finding is the easier detectability of Fv for the classroom compared to the other VAEs. A possible explanation can be identified in the larger percentage errors of the matched classroom's RIRs for most of the acoustical parameters reported in Tab. 2 ( $\ell_{end}$ ,  $T_S$  and  $C_{80}$ ). On the other hand, we found that participants were essentially unable to detect the impostor in the concert hall with Bin scenario. From an acoustic point of view, the directivity of LR is mitigated in a large environment, such as the concert hall. As a result, a static LR can be perceptually indistinguishable from a dynamic one in such an environment. For the living room, both Bin and Fv achieved low detection performances. This could be explained by the acoustic features of the living room whose short reverberation time (see Tab. 1) leaves a small amount of information for the listener to focus on. This is an advantage for the application of our findings since indoor environments, like the living room, are common for end-users.

As in the overall case, Bin is less detectable than Fv in complexities 3 and 4, and in classroom and concert hall VAEs. This was the expected behavior since the impostor condition  $F_d$  is an approximated version of  $B$ . However, we noticed that this pattern is reverted for complexity 2 and the living room. Though we found some statistically significant evidence of this pattern only for the logistic regression test (see Tab. 3) and not for  $d'$  values, this result represents an important finding for the use of an artificial reverberator to approximate LR. Nevertheless, for complexity 2, both Bin and Fv are quite detectable since is the most simple complexity level. As the complexity increases the task is harder, and Fv is less co-immersed than Bin which is almost undetectable.

In this paper, we proposed an experimental design in an AAV framework originally intended for the case study of LR. However, it can be exploited to evaluate the perceptual impact of an arbitrary rendering approach of virtual sound sources in an AAV scene. The main advantage of designing such experiences in AAV scenes is that the challenges involved in the creation of AAR scenes with real-world sound sources are avoided (e.g., tracking, real-time and responsive interactions, high-quality simulations, to name but a few [70]). As a result, more controlled and repeatable experimental settings can be designed within AAV. In particular, one can decide which aspects of the physical reality to capture and which interactions to handle in VR. Furthermore, any real environment can be simulated with a pair of headphones by means of the acquisition of spatial RIRs in such an environment.

Although the focus of this work is on the timbral characteristics of the late reverb, which is mainly diffuse, we used personalized HRTFs to improve the stimuli's ecological validity. However, the usefulness of personalized HRTFs is debatable and strongly influenced by context and task. Visual "capture" effects, where visual stimuli associated with a sound source affect its localization in space, have been extensively studied in experimental psychology [35]. Similar effects were recently observed in VR using generic HRTFs [9]. Thus, in future works, we plan to investigate the effect of different approaches and degrees of HRTF personalization, as well as to compare our acoustic simulation approach to other auralisation engines and simulators, such as Project Triton [58], in relation to the level of HRTF personalization.

## 6.1 Limitations

The encouraging results obtained from our experiment should be considered as a first step toward applying our methodology to the case study of LR. The low detectability of Bin and Fv reveals that the impostor's differences from the reference are barely perceived in our experiment. This effect is partially caused by the perceptual similarity achieved by the impostor conditions, but the experimental design may have a non negligible impact. As reported by many participants in a post-experiment interview, the simultaneous reproduction of the speakers prevents the listener to focus on the single speeches and their reverberation features. As a result, participants reported low confidence in their responses for most of the trials. These considerations suggest

that the level of difficulty of the task was too high. Thus, changes in the experiment's design will be crucial in future works. In particular, the scene with simultaneous speakers will be replaced with a conversation-like one. In this scene, the speakers will talk in sequence with partial or no overlapping, allowing the listener to focus on single speeches.

Another possible limitation of our experiment design is the full VR setting involved by AAV. The reference reverberation condition  $H$  was virtually rendered, as well as the impostor conditions  $B$  and  $F_d$ . The AAV design could intrinsically limit the applicability of the experiment's findings when the reference condition is not perceived as sufficiently authentic. For this reason, we performed all the needed operations to render the reverberation condition  $H$  as realistic as possible (HOA RIRs, individualized HRTFs, and HpTF compensation). Furthermore, we conducted the preliminary screening test to ascertain that the participants experienced ecologically rendered stimuli. Despite these precautions, the findings of our experiment cannot be directly extended to AAR settings. For this reason, in future works, we plan to design an experiment to transfer our results in AAR.

In the field of sonic interactions in VR/AR, the ultimate goal is the development of systems that account for the high variability in hearing sensitivity and characterization of the listeners [28]. Our recruited participants were far from being sufficiently representative of the potential user population, thus we refrain from drawing general conclusions. Despite this limitation, we included in the supplementary materials the experiment results grouped by the participants' characteristics (audio reverberation experience, age range, and sex). Even though the unbalanced and small sample sizes prevent a quantitative analysis, some qualitative observations can be made. In particular, we noticed that the higher the audio reverberation experience, the higher the impostor detectability, especially in the Fv scenario. In future works, we plan to expand upon this limitation for better generalization of the results.

Since the participants' screening was based on self-reported normal hearing conditions, as in several related works [16, 29, 45, 52, 54, 55], the generalization of the results could be further undermined. However, the preliminary screening test was designed to discard the participants experiencing not-ecologically valid stimuli. Despite that, a proper hearing sensitivity test would prevent biases caused by unknown hearing impairments [28, Ch. 1]. Thus, in future experiments, we plan to include more rigorous screening tests, e.g. speech-in-noise tests [38].

Our experiment failed in providing a worthwhile investigation of the identification task since the current protocol is inappropriate for meaningfully interpreting the obtained results. In future works, we plan to design an experiment protocol specific to the identification task.

## 7 CONCLUSION

In this paper, we described an experiment in an AAV setting to evaluate LR simplification without degradation on the co-immersion. We found that a static LR is perceived as co-immersed in a scene with other dynamic reference reverberation conditions. Conversely, when such a static LR is approximated with an artificial reverberator we found a significant impact on co-immersion, though limited. However, these findings are restricted to the evaluated scenario (simultaneous speakers) and within certain circumstances (complexity and VAE). Given the above discussion and the current experimental limitations, we plan to apply our methodology along four main directions to evaluate the influence on co-immersion of: (a) ERs simplification following previous related works [15, 31]. To this end, a possible artificial reverberator are Scattering Delay Networks (SDN) [18] which model ERs according to physical properties; (b) different listening environments, considering a wider set than the three VAEs used in the present study; (c) different conversational scenarios, considering concurrent talking [30], turn-taking dialogues, or partially overlapping speakers; (d) the effect of visual elements in modulating auditory perception and cognition, e.g. visual rendering matching the acoustics features (size, material, etc.).

## ACKNOWLEDGMENTS

This work is part of SONICOM, a project that has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101017743.

## REFERENCES

- [1] N. Agus, H. Anderson, J.-M. Chen, S. Lui, and D. Herremans. Minimally simple binaural room modeling using a single feedback delay network. *Journal of the Audio Engineering Society*, 66(10):791–807, Oct. 2018. doi: [10.17743/jaes.2018.0045](https://doi.org/10.17743/jaes.2018.0045) 4
- [2] V. R. Algazi, R. O. Duda, D. M. Thompson, and C. Avendano. The cipic hrtf database. In *Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (Cat. No. 01TH8575)*, pp. 99–102. IEEE, 2001. doi: [10.1109/ASPAA.2001.969552](https://doi.org/10.1109/ASPAA.2001.969552) 5
- [3] J. M. Arend, S. V. A. Garí, C. Schissler, F. Klein, and P. W. Robinson. Six-degrees-of-freedom parametric spatial audio based on one monaural room impulse response. *Journal of the Audio Engineering Society*, 69(7/8):557–575, 2021. doi: [10.17743/jaes.2021.0009](https://doi.org/10.17743/jaes.2021.0009) 2
- [4] F. Avanzini. *Procedural Modeling of Interactive Sound Sources in Virtual Reality*, pp. 49–76. Springer International Publishing, Cham, 2023. doi: [10.1007/978-3-031-04021-4\\_2](https://doi.org/10.1007/978-3-031-04021-4_2) 8
- [5] H. Bahu and D. Romblo. Optimization and prediction of the spherical and ellipsoidal itd model parameters using offset ears. In *Audio Engineering Society Conference: 2018 AES International Conference on Spatial Reproduction-Aesthetics and Science*. Audio Engineering Society, 2018. 4
- [6] W. Bailey and B. Fazenda. The effect of visual cues and binaural rendering method on plausibility in virtual environments. In *Audio Engineering Society Convention 144*. Audio Engineering Society, may 2018. 2, 8
- [7] D. R. Begault and L. J. Trejo. *3-D sound for virtual reality and multimedia*. 09 2000. 1
- [8] D. R. Begault, E. M. Wenzel, and M. R. Anderson. Direct comparison of the impact of head tracking, reverberation, and individualized head-related transfer functions on the spatial perception of a virtual speech source. *Journal of the Audio Engineering Society*, 49(10):904–916, february 2001. 4
- [9] C. C. Berger, M. Gonzalez-Franco, A. Tajadura-Jiménez, D. Florencio, and Z. Zhang. Generic hrtfs may be good enough in virtual reality, improving source localization through cross-modal plasticity. *Frontiers in Neuroscience*, 12, 2018. doi: [10.3389/fnins.2018.00021](https://doi.org/10.3389/fnins.2018.00021) 9
- [10] I. Bergström, S. Azevedo, P. Papiotis, N. Saldanha, and M. Slater. The plausibility of a string quartet performance in virtual reality. *IEEE transactions on visualization and computer graphics*, 23(4):1352–1359, 2017. doi: [10.1109/TVCG.2017.2657138](https://doi.org/10.1109/TVCG.2017.2657138) 2
- [11] B. Bernschütz. A spherical far field hrir/hrtf compilation of the neumann ku 100. In *Proceedings of the 40th Italian (AIA) annual conference on acoustics and the 39th German annual conference on acoustics (DAGA) conference on acoustics*, p. 29. German Acoustical Society (DEGA) Berlin, 2013. 2, 3, 4
- [12] V. Best, R. Baumgartner, M. Lavandier, P. Majdak, and N. Kopčo. Sound Externalization: A Review of Recent Research. *Trends in Hearing*, 24, Jan. 2020. Publisher: SAGE Publications Inc. doi: [10.1177/2331216520948390](https://doi.org/10.1177/2331216520948390) 6
- [13] J. Blauert. *Spatial Hearing: The Psychophysics of Human Sound Localization*. The MIT Press, 10 1996. doi: [10.7551/mitpress/6391.001.0001](https://doi.org/10.7551/mitpress/6391.001.0001) 2
- [14] R. Bona, D. Fantini, G. Presti, M. Tiraboschi, J. I. Engel Alonso-Martinez, and F. Avanzini. Automatic parameters tuning of late reverberation algorithms for audio augmented reality. In *Proceedings of the 17th International Audio Mostly Conference, AM '22*, pp. 36–43. Association for Computing Machinery, New York, NY, USA, 2022. doi: [10.1145/3561212.3561236](https://doi.org/10.1145/3561212.3561236) 2, 3, 4, 8
- [15] F. Brinkmann, H. Gamper, N. Raghuvanshi, and I. Tashev. Towards encoding perceptually salient early reflections for parametric spatial audio rendering. In *Audio Engineering Society Convention 148*. Audio Engineering Society, may 2020. 9
- [16] F. Brinkmann, A. Lindau, and S. Weinzierl. On the authenticity of individual dynamic binaural synthesis. *The Journal of the Acoustical Society of America*, 142(4):1784–1795, 2017. doi: [10.1121/1.5005606](https://doi.org/10.1121/1.5005606) 2, 9
- [17] A. D. Brown, G. C. Stecker, and D. J. Tollin. The precedence effect in sound localization. *Journal of the Association for Research in Otolaryngology*, 16:1–28, 2015. doi: [10.1007/s10162-014-0496-2](https://doi.org/10.1007/s10162-014-0496-2) 2
- [18] E. De Sena, H. Hacıhabiboğlu, Z. Cvetković, and J. O. Smith. Efficient synthesis of room acoustics via scattering delay networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(9):1478–1492, 2015. doi: [10.1109/TASLP.2015.2438547](https://doi.org/10.1109/TASLP.2015.2438547) 9
- [19] J. Eaton, N. D. Gaubitch, A. H. Moore, and P. A. Naylor. Estimation of room acoustic parameters: The ACE challenge. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(10):1681–1693, 2016. doi: [10.1109/TASLP.2016.2577502](https://doi.org/10.1109/TASLP.2016.2577502) 5
- [20] K. Enge, M. Frank, and R. Höldrich. Listening experiment on the plausibility of acoustic modeling in virtual reality. In *Fortschritte der Akustik–DAGA*, vol. 46, pp. 13–16, 2020. 2
- [21] I. Engel, C. Henry, S. V. Amengual Garí, P. W. Robinson, and L. Picinali. Perceptual implications of different ambisonics-based methods for binaural reverberation. *The Journal of the Acoustical Society of America*, 149(2):895–910, 2021. doi: [10.1121/10.0003437](https://doi.org/10.1121/10.0003437) 2
- [22] J. Engel, L. H. Hantrakul, C. Gu, and A. Roberts. Ddsp: Differentiable digital signal processing. In *International Conference on Learning Representations*, 2020. 3
- [23] D. Fantini, G. Presti, M. Geronazzo, R. Bona, A. G. Privitera, and F. Avanzini. Project’s repository for: Co-immersion in Audio Augmented Virtuality: the Case Study of a Static and Approximated Late Reverberation Algorithm, June 2023. doi: [10.5281/zenodo.8026357](https://doi.org/10.5281/zenodo.8026357) 2
- [24] A. Farina. Auralization software for the evaluation of a pyramid tracing code: results of subjective listening tests. In *ICA95 (International Conference on Acoustics), Trondheim (Norway)*, pp. 26–30, 1995. 4, 5
- [25] A. Farina. Aurora software. [http://pcf.arina.eng.unipr.it/aurora\\_xp/index.htm](http://pcf.arina.eng.unipr.it/aurora_xp/index.htm), 2010. version 4.3. 4, 5
- [26] I. O. for Standardization. *ISO 3382-1: International Standard ISO/DIS 3382-1: Acoustics – Measurement of room acoustic parameters – Part 1: Performance spaces*. International Organization for Standardization, 2009. 3, 5
- [27] M. Geronazzo, E. Peruch, F. Prandoni, and F. Avanzini. Applying a single-notch metric to image-guided head-related transfer function selection for improved vertical localization. *Journal of the Audio Engineering Society*, 67(6):414–428, June 2019. doi: [10.17743/jaes.2019.0010](https://doi.org/10.17743/jaes.2019.0010) 5
- [28] M. Geronazzo and S. Serafin, eds. *Sonic Interactions in Virtual Environments*. Human-Computer Interaction Series. Springer International Publishing, Cham, 1 ed., 2023. doi: [10.1007/978-3-031-04021-4\\_1](https://doi.org/10.1007/978-3-031-04021-4_1) 9
- [29] M. Geronazzo, J. Y. Tissieres, and S. Serafin. A minimal personalization of dynamic binaural synthesis with mixed structural modeling and scattering delay networks. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 411–415. IEEE, 2020. doi: [10.1109/ICASSP40776.2020.9053873](https://doi.org/10.1109/ICASSP40776.2020.9053873) 4, 9
- [30] M. Gonzalez-Franco, A. Maselli, D. Florencio, N. Smolyanskiy, and Z. Zhang. Concurrent talking in immersive virtual reality: on the dominance of visual speech cues. *Scientific Reports*, 7(1):3817, Jun 2017. doi: [10.1038/s41598-017-04201-x](https://doi.org/10.1038/s41598-017-04201-x) 9
- [31] H. Hacıhabiboğlu and F. Murtagh. Perceptual simplification for model-based binaural room auralisation. *Applied Acoustics*, 69(8):715–727, 2008. doi: [10.1016/j.apacoust.2007.02.006](https://doi.org/10.1016/j.apacoust.2007.02.006) 9
- [32] C. C. Hak, R. H. Wenmaekers, and L. Van Luxemburg. Measuring room impulse responses: Impact of the decay range on derived room acoustic parameters. *Acta Acustica united with Acustica*, 98(6):907–915, 2012. doi: [10.3813/AAA.918574](https://doi.org/10.3813/AAA.918574) 3
- [33] M. J. Hautus, N. A. Macmillan, and C. D. Creelman. *Detection theory: A user’s guide*. Routledge, 2021. 7
- [34] T. Hidaka, L. L. Beranek, and T. Okano. Interaural cross-correlation, lateral fraction, and low-and high-frequency sound levels as measures of acoustical quality in concert halls. *The Journal of the Acoustical Society of America*, 98(2):988–1007, 1995. doi: [10.1121/1.414451](https://doi.org/10.1121/1.414451) 4
- [35] C. E. Jack and W. R. Thurlow. Effects of degree of visual association and angle of displacement on the “ventriloquism” effect. *Perceptual and Motor Skills*, 37(3):967–979, 1973. PMID: 4764534. doi: [10.1177/003151257303700360](https://doi.org/10.1177/003151257303700360) 9
- [36] J. Jerald. *The VR Book: Human-Centered Design for Virtual Reality*. Association for Computing Machinery and Morgan & Claypool, 2015. doi: [10.1145/2792790](https://doi.org/10.1145/2792790) 1
- [37] JUCE. Freeverb. [https://github.com/juce-framework/JUCE/blob/master/modules/juce\\_audio\\_basics/utilities/juce\\_Reverb.h](https://github.com/juce-framework/JUCE/blob/master/modules/juce_audio_basics/utilities/juce_Reverb.h), 2022. 4
- [38] M. C. Killion, P. A. Niquette, G. I. Gudmundsen, L. J. Revit, and S. Banerjee. Development of a quick speech-in-noise test for measuring signal-to-noise ratio loss in normal-hearing and hearing-impaired listeners. *The Journal of the Acoustical Society of America*, 116(4):2395–2405, 10 2004. doi: [10.1121/1.1784440](https://doi.org/10.1121/1.1784440) 9
- [39] M. Kleiner, B.-I. Dalenbäck, and P. Svensson. Auralization – An overview. *Journal of the Audio Engineering Society*, 41(11):861–875, 1993. 1
- [40] M.-V. Laitinen and V. Pulkki. Binaural reproduction for directional audio coding. In *2009 IEEE Workshop on Applications of Signal Processing*

- to *Audio and Acoustics*, pp. 337–340. IEEE, 2009. doi: 10.1109/ASPA.2009.5346545 3
- [41] H. Lee and D. Johnson. 3D Microphone Array Recording Comparison (3D- MARCo), Oct. 2019. doi: 10.5281/zenodo.3477602 1, 3
- [42] H. Lee and D. Johnson. An open-access database of 3d microphone array recordings. In *Audio Engineering Society Convention 147*. Audio Engineering Society, october 2019. 1, 3
- [43] S. Lee, H.-S. Choi, and K. Lee. Differentiable artificial reverberation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:2541–2556, 2022. doi: 10.1109/TASLP.2022.3193298 3
- [44] A. Lindau, L. Kosanke, and S. Weinzierl. Perceptual evaluation of model- and signal-based predictors of the mixing time in binaural room impulse responses. *Journal of the Audio Engineering Society*, 60(11):887–898, november 2012. 2
- [45] A. Lindau and S. Weinzierl. Assessing the plausibility of virtual acoustic environments. *Acta Acustica united with Acustica*, 98(5):804–810, 2012. doi: 10.3813/AAA.918562 2, 8, 9
- [46] T. Lokki and J. Pätynen. *Auditory Spatial Impression in Concert Halls*, pp. 173–202. Springer International Publishing, Cham, 2020. doi: 10.1007/978-3-030-00386-9\_7 2
- [47] A. H. Marshall and M. Barron. Spatial responsiveness in concert halls and the origins of spatial impression. *Applied Acoustics*, 62(2):91–108, 2001. doi: 10.1016/S0003-682X(00)00050-5 2
- [48] L. McCormack and A. Politis. Sparta & compass: Real-time implementations of linear and parametric spatial audio reproduction and processing methods. In *Audio Engineering Society Conference: 2019 AES International Conference on Immersive and Interactive Audio*. Audio Engineering Society, march 2019. 4
- [49] P. Milgram and F. Kishino. A taxonomy of mixed reality visual displays. *IEICE TRANSACTIONS on Information and Systems*, 77(12):1321–1329, 1994. 1
- [50] H. Møller, M. F. Sørensen, C. B. Jensen, and D. Hammershøi. Binaural technique: Do we need individual recordings? *Journal of the Audio Engineering Society*, 44(6):451–469, june 1996. 4
- [51] K. R. Murphy and B. Myers. Testing the hypothesis that treatments have negligible effects: Minimum-effect tests in the general linear model. *Journal of Applied Psychology*, 84(2):234–248, 1999. doi: 10.1037/0021-9010.84.2.234 8
- [52] A. Neidhardt, A. Ignatious-Tommy, and A. D. Pereppadan. Plausibility of an interactive approaching motion towards a virtual sound source based on simplified brir sets. In *Audio Engineering Society Convention 144*. Audio Engineering Society, may 2018. 2, 9
- [53] A. Neidhardt, C. Schneiderwind, and F. Klein. Perceptual matching of room acoustics for auditory augmented reality in small rooms-literature review and theoretical framework. *Trends in Hearing*, 26:23312165221092919, 2022. doi: 10.1177/23312165221092919 2
- [54] A. Neidhardt and A. M. Zerlik. The availability of a hidden real reference affects the plausibility of position-dynamic auditory ar. *Frontiers in Virtual Reality*, 2, 2021. doi: 10.3389/frvir.2021.678875 2, 9
- [55] J. Oberem, B. Masiero, and J. Fels. Experiments on authenticity and plausibility of binaural reproduction via headphones employing different recording methods. *Applied Acoustics*, 114:71–78, 2016. doi: 10.1016/j.apacoust.2016.07.009 2, 9
- [56] O. Olgun and H. Hacıhabiboğlu. METU SPARG Eigenmike em32 Acoustic Impulse Response Dataset v0.1.0, Apr. 2019. doi: 10.5281/zenodo.2635758 1, 3
- [57] C. Pike, F. Melchior, and T. Tew. Assessing the plausibility of non-individualised dynamic binaural synthesis in a small room. In *Audio Engineering Society Conference: 55th International Conference: Spatial Audio*. Audio Engineering Society, august 2014. 2, 8
- [58] N. Raghuvanshi and J. Snyder. Parametric directional coding for precomputed sound propagation. *ACM Trans. Graph.*, 37(4), jul 2018. doi: 10.1145/3197517.3201339 9
- [59] F. Rumsey. Evaluating AVAR: Goodbye quality, hello plausibility? *Journal of the Audio Engineering Society*, 66(12):1126–1130, 2018. 1
- [60] V. Satopaa, J. Albrecht, D. Irwin, and B. Raghavan. Finding a "kneedle" in a haystack: Detecting knee points in system behavior. In *31st international conference on distributed computing systems workshops*, pp. 166–171. IEEE, 2011. doi: 10.1109/ICDCSW.2011.20 3
- [61] J. Schütze, C. Kirsch, K. C. Wagener, B. Kollmeier, and S. D. Ewert. Living room environment, Sept. 2021. Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Projektnummer 352015383 – SFB 1330, Project C4 and C5. doi: 10.5281/zenodo.5747753 1, 3
- [62] M. Slater, B. Spanlang, and D. Corominas. Simulating virtual environments within virtual environments as the basis for a psychophysics of presence. *ACM Transactions on Graphics*, 29(4):92:1–92:9, July 2010. doi: 10.1145/1778765.1778829 6
- [63] J. O. Smith. *Physical Audio Signal Processing*. W3K Publishing, 2010. 2
- [64] J. Snoek, H. Larochelle, and R. P. Adams. Practical bayesian optimization of machine learning algorithms. In *Advances in Neural Information Processing Systems*, vol. 25, 2012. 3
- [65] H. Stanislaw and N. Todorov. Calculation of signal detection theory measures. *Behavior research methods, instruments, & computers*, 31(1):137–149, 1999. doi: 10.3758/BF03207704 8
- [66] G. C. Stecker, T. M. Moore, M. Folkerts, D. Zotkin, and R. Duraiswami. Toward objective measures of auditory co-immersion in virtual and augmented reality. In *Audio Engineering Society Conference: 2018 AES International Conference on Audio for Virtual and Augmented Reality*. Audio Engineering Society, august 2018. 1, 2
- [67] E. M. Wenzel, M. Arruda, D. J. Kistler, and F. L. Wightman. Localization using nonindividualized head-related transfer functions. *The Journal of the Acoustical Society of America*, 94(1):111–123, July 1993. doi: 10.1121/1.407089 4
- [68] C. K. Williams and C. E. Rasmussen. *Gaussian processes for machine learning*, vol. 2. MIT press Cambridge, MA, 2006. 3
- [69] S. A. Wirlner, N. Meyer-Kahlen, and S. J. Schlecht. Towards transfer-plausibility for evaluating mixed reality audio in complex scenes. In *Audio Engineering Society Conference: 2020 AES International Conference on Audio for Virtual and Augmented Reality*. Audio Engineering Society, august 2020. 2
- [70] J. Yang, A. Barde, and M. Billinghurst. Audio augmented reality: A systematic review of technologies, applications, and future research directions. *Journal of the Audio Engineering Society*, 70(10):788–809, october 2022. doi: 10.17743/jaes.2022.0048 1, 9
- [71] F. Zotter and M. Frank. *Ambisonics – A Practical 3D Audio Theory for Recording, Studio Production, Sound Reinforcement, and Virtual Reality*. Springer Cham, 2019. 1