Audio Engineering Society

Conference Paper 4

Presented at the 6th International Conference on Audio for Games
2024 April 27–29, Tokyo, Japan

This paper was peer-reviewed as a complete manuscript for presentation at this conference. This paper is available in the AES E-Library (http://www.aes.org/e-lib), all rights reserved. Reproduction of this paper, or any portion thereof, is not permitted without direct permission from the Journal of the Audio Engineering Society.

CGI Scenes for Interactive Audio Research and Development: Cave, Cinema, and Mansion

Thomas Robotham[1], Daniela Rebmann[2], Dominik O. Fintineanu-Anghelescu[2], Alexander Raake[3], and Emanuël A. P. Habets[1]

[1]*International Audio Laboratories Erlangen, Erlangen, Germany**
[2]*Fraunhofer-Institut für Integrierte Schaltungen IIS, Erlangen, Germany*
[3]*Audiovisual Technology Group, TU-Ilmenau, Ilmenau, Germany*

Correspondence should be addressed to Thomas Robotham (thomas.robotham@audiolabs-erlangen.de)

ABSTRACT

Audio rendering engines are a cornerstone in offering a plausible and immersive experience for interactive virtual environments (IVEs). For virtual reality IVEs, a culmination of visuals, audio, interactive, and behavioral cues blend to form a user's perception and cognition. However, implementing such IVEs incurs additional costs and resources beyond the scope of many labs. This contribution describes a set of three open-source computer-generated imagery interactive audiovisual scenes, including geometric, material, lighting, and post-processing implementation for relevant audio and visual cues. In addition, each IVE poses an audio-relevant task for users to perform throughout the environment, invoking cognitive processes for further psychological and behavioral research. The results of a small-scale case study are presented, which demonstrate the IVE design's impact on user behavior along with scene profiling of selected acoustic attributes. The scene profiling highlights that different acoustic auralization attributes for IVEs may be needed as a combination of both the IVE's physical design and the user task.

1 Introduction

One of the core components of virtual reality (VR) experiences lies in the interactive virtual environment (IVE), a culmination of visual graphics, immersive high-fidelity audio, and user interactivity. Combined with a narrative, compelling task, or game-like objective, a user will explore the IVE based on top-down motivations and bottom-up sensory input. As a multi-modal experience, the inter-dependency and coherence between these factors become ever more critical, to the degree where the quality of an individual aspect may impact the perceived quality and cognitive impact of another aspect. Consequently, although modular quality testing of a particular audio or visual feature is an essential component of any IVE system, meaningful perceptual, cognitive, and behavioral insights can be gained using high-quality multi-modal IVEs and involving more complex, interconnected tasks. Such environments serve as invaluable tools for researching the underlying quality of experience of the complete VR system, and individual technological choices, for

*A joint institution of the Friedrich-Alexander-UniversitäFriedrich-Alexander-Universitt Erlangen-NüErlangen-Nrnberg (FAU) and Fraunhofer Institute for Integrated Circuits (IIS).

example, regarding the audio modality.

1.1 Multi-modal IVEs

Developing audio for VR IVEs is a multifaceted challenge. Such challenges include capturing source material (anechoic vs. wet signals), source and receiver properties (e.g., position, orientation, directivity patterns / head-related transfer functions [1]), implementing render functionality (e.g., reverberation, occlusion, diffraction, [2, 3, 4, 5]), crafting additional acoustic environment (e.g., absorption/scattering coefficients and geometry [6]), to real-time audio source interactivity. Consequently, much effort has gone into creating rendering engines that deliver high-fidelity audio that supports an immersive experience.

In contrast, few studies have focused on using high-quality audiovisual interactive virtual reality scenes. Par et al. [1] crafted three high-quality scenes for the purposes of hearing science. The scenes provide a complex set of acoustic geometries in real-life-like settings and high-quality visuals for ecologically valid multi-modal research with ground truth recordings. While targeting real-life-like presentation, the scenes do not focus on delivering interactive elements often seen in virtual reality experiences. In a global effort, multiple IVEs have been crafted during the development of a new MPEG audio standard (MPEG-I) [7]. However, the scenes are not available to external parties. Llorca-Bofí and Vorländer [6] provide an in-depth architectural framework for virtual reality acoustics and sound perception. Here, the main focus is placed on the material and geometric composition of the IVE to provide a high-fidelity rendering of a classroom space. Due to the architectural context of the framework, little is described in terms of audiovisual interactivity.

To facilitate research on auditory perception and cognition in multi-modal VR IVEs across communities, open-source scenes incorporating realistic features and properties for audio rendering and high-quality visuals that also allow for user-object interaction would be a valuable addition.

1.2 Engagement and Experience

While sensory fidelity takes center stage at a systems level, the key element of an IVE is the user. In game design, a compelling narrative or purpose within the scene offers more than just the presentation of multi-modal stimuli [8]. Here, the concept of '*experiential fidelity*' focuses on the user experience based on their perception of a narrative or story exposed through word building and the alignment of expectations of what users are presented [9, 10]. More recently, avenues of audio evaluation for VR employ umbrella constructs such as immersion [11], plausibility [12], and quality of experience (QoE) [13]. These terms share the same principle as experiential fidelity, leveraging a user's inner-reference (of a particular construct) to formulate a judgment based on their expectations, some of which share overlapping components. For example, comparable to '*experiential fidelity*', '*internal plausibility*' defined by Hofer et al. [14] is the consistency of stimuli against the user's expectations raised as a consequence of the presented genre or world.

In interactive audio evaluation scenarios, utilizing perspectives from game design and experiential fidelity could complement evolving metrics such as immersion, QoE, and plausibility. Offering more than just the presentation of (high quality) multi-modal content may provide subjects a reason to engage and experience the world. Of course, an explicit audio quality evaluation task is also a reason to explore and engage. However, this task offers little in terms of immersion. Indeed, the opposite is more likely (and desirable), whereby subjects must actively maintain an observational state of mind and not forget the primary judgment task. By designing a narrative or game-like task within the VR IVE for subjects to experience, we may (a) elicit behavior and engagement more representative of realistic end-user scenarios and (b) complement constructs that aim to quantify an experience supported by audiovisual and interactive systems.

Recent studies are beginning to adopt this approach. For example, Letter et al. [13] adopt a diverse set of test stimuli (e.g., [15]) that include explicit story narratives to ascertain subjects' personal constructs towards QoE. Poirier-Quinot and Katz [16] assess performance metrics of shooter-style scenarios commonplace among gaming. Similar to [17], a recent study by Garí et al. combined a labyrinth and way-finding task to evaluate audio rendering on navigation performance [18]. For future studies, a set of open source, carefully designed scenes, with audio playing an active role in the task (e.g., driving attention and encouraging exploration [19]) would add considerable added value to behavioral and cognitive research.

Fig. 1: Renders of the developed IVEs. From left to right: CAVE, CINEMA, and MANSION. (Note: Brightness and contrast have been increased here for ease of illustration.)

1.3 Contribution Statement

This article describes a set of open-source IVEs created for multi-disciplinary research and development. The scenes primarily focus on auditory perception and cognition research and algorithmic development for VR but may also extend to further fields within behavioral sciences and psychophysiological research. This article describes the visual modeling, interactive audio design process, and considerations for more advanced acoustic auralization. Moreover, the article serves as a template for those wishing to develop further IVEs. An additional feature within the scenes is the potential for user tasks that allow for more holistic and indirect evaluations for perceptual, cognitive, and behavioral analysis. Each user task is designed with a focus on audio, ensuring that presence, rendering quality, and acoustic attributes are inherently associated with the task at hand.

The release of the VR IVEs is accompanied by evaluation data from a small-scale subjective scene analysis. As each scene is designed to focus on different acoustic elements and assets, the subjective data demonstrate the range of exploration possibilities across the scenes through subject movement. All scenes and supporting technical information are available open source under the CC BY-NC 4.0 license at https://qoevave.github.io/database/.

2 IVE Framework

The Unity game engine was used to develop the IVEs[1]. Unity was chosen due to the accessible scripting framework and commonplace usage in many research studies (e.g., [11, 18, 20, 21, 22], and use within standardization activities [7]. Specifically, version 2021.3 LTS (long-term support) High Definition Render Pipeline (HDRP) version 12.1.7[2] was used. For virtual reality support, OpenXR (v1.7.0)[3] is enabled by default within the XR Plug-in management (v4.2.1)[4], which can be switched out for Oculus support when using Quest devices.

2.1 General Design

Considering previous studies on audio-relevant level design [19, 23], and aspects of gamification and VR, we generated multiple IVE concepts guided by three general design concepts. (i) The scenes should possess various acoustic attributes and diverse audio signals. As audio sources, both audio-only and audiovisual objects should be present in the scene. (ii) The scenes should be designed with the possibility of a game-like user objective. The audio implementation for this gaming task should have an influential role. (iii) The scenes must cater to six-degrees-of-freedom (6-DoF) movement, possess interactive objects, and several assets or visual points of interest to foster user exploration, behavior, and engagement. After several design iterations, the scenes **CAVE**, **CINEMA**, and **MANSION** were selected. The semantic setting of the selected scenes, level design concepts, and asset lists (i.e., audiovisual objects) promise diverse interactions, user-task potential, audio source types and acoustic attributes, and visual aesthetics. Figure 1 shows the final scene designs using still image renders at certain positions.

The CAVE (Figure 1 - left) concept lends itself to user exploration driven by audio cues. To emphasize this,

[1] https://unity.com/

[2] https://unity.com/srp/High-Definition-Render-Pipeline

[3] https://docs.unity3d.com/Packages/com.unity.xr.openxr@1.7

[4] https://docs.unity3d.com/Manual/com.unity.xr.management.html

the lighting within the scene is relatively dark, with some sunlight leaking through gaps in the cave structure. To help users navigate, an interactive lantern is provided that emits a soft directional spot light. Multiple audio sources can be placed in different positions within the scene, including an upper and lower floor, which can be accessed from two sides of the map. Visual points of interest include a river with growing mushrooms, a larger lagoon with many stalagmites and stalactites, and a light opening with flora.

The CINEMA (Figure 1 - middle) allows for a scene of high ecological validity with the likelihood that most users have experienced a cinema before. The cinema space was designed as a virtual counterpart to a real-life cinema located at Fraunhofer IIS, allowing virtual acoustics to be compared against real-life audio. Moreover, the setting offers a stimuli-within-stimuli dimension, meaning that this scene serves as stimuli and that further audiovisual stimuli can be reproduced within the IVE. Visual points of interest include human avatars and the cinema media playback.

The MANSION (Figure 1 - right) is designed such that multiple visual points of interest are presented to the user (e.g., paintings, a chess set, a piano, etc.). Many of these points of interest also possess audio sources that can be triggered to playback at certain events based on user interaction. Unlike the previous two scenes, the MANSION features many materials that become highly relevant when considering interactive objects reacting to surface impacts.

The IVEs should also offer various acoustic properties, as a point of general design layout. While the auralization itself is mainly dependent on the employed audio renderer, the potential for specific acoustic attributes should be promoted by each IVE's design. Consequently, the scenes have been constructed to offer a range of attributes relevant to audio rendering, such as late reverberation, occlusion, and source extent.

2.2 Visual

To construct the 3D scene models, *Blender* and *Cinema 4D* modeling software was used. Here, individual models such as walls, structures, and objects were designed. The 3D models were then UV unwrapped and imported into *3D Substance Painter*, where textures can be applied corresponding to the UV layout. Afterward, all models and textures were imported into

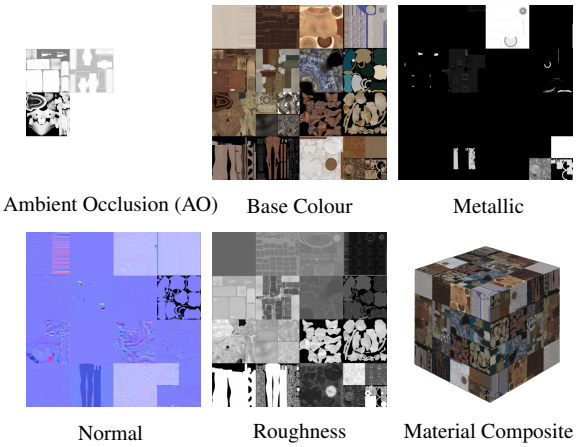

Fig. 2: Composition of texture atlases for visual materials.

Unity as assets and subsequently used to construct the IVE. The models and textures were either specifically designed for the three scenes or made using CC-0 licensed files. As many different materials can lead to resource-intensive operations, most textures have been collected into 'texture atlases' shown in Fig. 2. These allow the use of only one material for several scene objects, while maintaining different material properties like color, roughness, or metallicness.

Unity's high-definition render pipeline (HDRP) was used to allow for high-fidelity graphics and additional options, such as volumetric fog/lighting rays utilized in the CAVE and MANSION scenes. As lighting computation in VR IVEs is resource intensive, the number of real-time lights is kept low, and lighting information has been baked into light maps, which store data for all static objects. Dynamic objects are lit using so-called 'light probes'. Light probes measure light at different positions in the scene. This information can then be used at run-time to approximate the lighting effects of the nearest objects. Both light maps and light probe data are calculated prior to the application running. In addition to the local light sources, High Dynamic Range Images are used for global lighting. These images also serve as a visible background for the CAVE and MANSION. Lights that change during runtime and thus cannot be baked include an interactive lantern in the CAVE and a screen light in the CINEMA, which differs depending on virtual cinema screen pixels. This screen light uses a pre-baked animation to change color. The CINEMA also includes a button that can be pressed to turn on several ceiling lights to make the scene brighter. For all three scenes, post-processing

effects can be freely adjusted.

A considerable challenge of high-fidelity VR IVEs is to achieve an appropriate rendering performance. To this end, most objects use 'levels of detail' (commonly referred to as 'LODs'). Depending on the distance of an object to the user position, rendered meshes are swapped out for lower-detailed versions. Objects that are further away have fewer polygons and thus require less computing power. Moreover, 3D objects in Unity can be marked as static, meaning they never change during run-time. Marking objects as static allows for the meshes to be combined and reduces the resources needed to render them.

2.3 Audio

The audio was created through a combination of CC-0 license files and curated recordings captured via a Zoom H5 recorder. A catalog of implemented audio files can be found in the accompanying online resources (see Sec. 4). For the interactive sound design, two types of audio are implemented: *continuous*, and *event-based audio*. *Continuous* audio initializes on scene start and continues to play throughout the experience. *Event-based* audio reacts to subjects' interactions (i.e., an impact sound when throwing an interactive object). Thus, in scenes where multiple different surface materials are present, different impact sounds will be heard.

To this end, multiple impact sounds were recorded for various interactable objects. For example, the interactable chess pieces in the MANSION IVE have impact sounds for the surfaces of carpet, vases (i.e., porcelain), glass, metal, wood, flooring, textile (i.e., sofa/cushions), stone, etc. For each surface impact, multiple audio files are present to reduce the repetition present when triggering the same audio file. Moreover, the user may throw intractable objects at varying speeds, thus requiring multiple impact sounds at different levels to reflect impact velocity. Therefore, the correct sound file can be selected via event detection within the Unity game engine to decipher the material collision and impact velocity. Currently, the scripts provided within the scenes differentiate between low- and high-velocity magnitudes at a threshold of 2 m/s. Overall, the total audio files for each IVE is 380 for the MANSION, 117 for the CINEMA, and 169 for the CAVE.

Object-based audio is the predominant workflow for immersive audio rendering in all scenes, with all audio files provided in mono at 48 kHz sample rate. For the CINEMA IVE, additional channel-based or scene-based content may be included as virtual loudspeakers within the cinema (i.e., a 7.1+4 channel-based production can be implemented as objects at the correct loudspeaker positions). Event-based audio files are trimmed at zero-crossing to the transient onset, ensuring audio playback begins precisely at the time of impact. The MetaXR Audio SDK[5] was employed to provide real-time interactive audio rendering. The SDK extends the default Unity audio spatializer functionality to include near-field rendering, room acoustic simulation, Ambisonics support, and direct-to-reverberation ratio cues based on reflections and distance. With this, a basic plug-and-play audio solution is provided at the cost of more advanced rendering features such as occlusion, diffraction, and complex acoustic scene geometry.

For developers and research requiring more advanced or in-house solutions, the MetaXR audio SDK can be exchanged via the audio spatializer options in the project settings. To accommodate more complex acoustic auralization techniques, additional handcrafted meshes, as shown in Fig. 3, are provided. Each mesh is stored in the respective Unity scene as a separate object. All provided meshes are so-called solid meshes (i.e., with a governing thickness). For audio rendering, this ensures the direct sound only interacts with one side of a mesh.

2.4 Interactions, Adaptability, and Tasks

The three scenes have been constructed with the following potential user tasks focusing on auditory perception and audio rendering features. Inspired by [17, 18], the main task for the CAVE is a way-finding challenge in a visually dark setting. While visual information is limited by a field of view, audio information is a consistent stream of 360° spatial information. Therefore, users are given an interactive lantern to help them visually but must rely on audio cues to navigate effectively. The way-finding target is a walkie-talkie that can spawn at five locations (see Table 1). Consequently, the capabilities of the audio rendering relevant to the scene may impact subjects' task performance and behavior. The CINEMA IVE provides a realistic opportunity to evaluate audiovisual content already mixed via a sound engineer. One potential user task is for subjects to

[5] https://developer.oculus.com/documentation/unity/meta-xr-audio-sdk-unity/

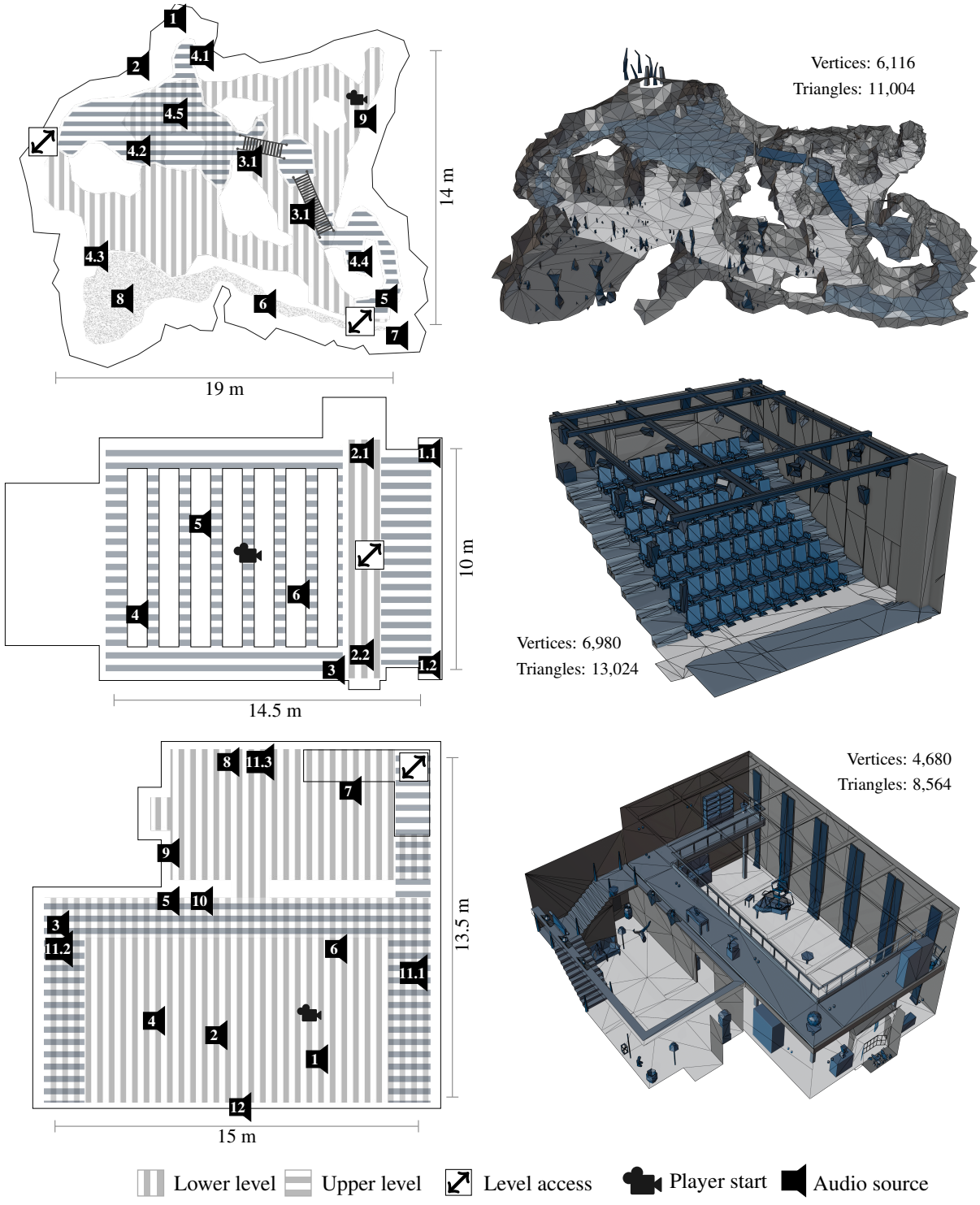

Fig. 3: Left: Illustration of scene layout and level design of upper and lower floors, starting position, and position of audio object sources within the scene corresponding with data in Table 1. Right: Provided meshes for the corresponding scenes. Scenes from top to bottom, the CAVE, CINEMA, and MANSION. Note: Selected wall and ceiling meshes have been removed for illustration purposes.

Table 1: Audio source ID#, descriptions and initial object positions within the three scenes (position transform is given Unity coordinate system). The total number of audio files within each scene is indicated next to the scene name.

	ID#	Audio source description	Initial position (x, y, z)
CAVE (169 files)	1	Bats ↳ Static chirps and alarmed sounds when triggered.	(-14.82, 4.76, -11.45)
	2	Outside environment	(-22.86, 6.75, -16.63)
	3.1	Creaking wood	(-2.79, 3.25, -12.2)
	3.2	Creaking wood	(4.94, 2.62, -12.8)
	4.1	Walkie-talkie #1	(-13.3, 3.21, -12.69)
	4.2	Walkie-talkie #2	(-11.0, 3.43, -22.43)
	4.3	Walkie-talkie #3	(-7.35, 0.09, -29.29)
	4.4	Walkie-talkie #4	(11.73, 2.69, -13.54)
	4.5	Walkie-talkie #5	(11.73, 2.69, -13.54)
	5	Wind howling	(12.26, 6.99, -13.82)
	6	Water stream ↳ Includes multiple source positions of water located along the stream.	(6.75, 0.18, -22.4)
	7	Stream waterfall	(18.73, 2.48, -15.62)
	8	Water droplets ↳ Multiple audio positions based on water droplet particle system	(-3.26, 0, -31.79)
	9	Lantern	(0.66, 0.38, -0.57)
CINEMA (117 files)	1.1	Curtain motor #1	(-4.73, 4.82, 8.23)
	1.2	Curtain motor #2	(5.40, 4.82, 8.23)
	2.1	Loudspeaker (L)	(-4.90, 1.63, 5.91)
	2.2	Loudspeaker (R)	(5.45, 1.63, 5.91)
	3	Lights toggle click	(5.72, -0.73, 4.24)
	4	Human sipping drink	(2.98, 1.19, -4.99)
	5	Phone call	(-0.96, 1.0, -2.15)
	6	Human eating	(2.21, -0.4, 2.16)
MANSION (380 files)	1	Piano	(3, 1, -2.35)
	2	Chess Pieces ↳ Chess board of 36 intractable pieces.	(-1.38, 0.77, -1.26)
	3	Plates moving	(3.97, 0.8, 1.68)
	4	Furniture moving	(-3.8, 0.1, -0.6)
	5	Knight armour	(-2.82, 1, 4.1)
	6	Marble statue	(-8.25, 1.21, 2.197)
	7	Crate of bottles ↳ Holds six interactive bottles.	(5.25, 0.21, 8.96)
	8	Umbrella holder ↳ Holds two intractable umbrellas, and three intractable walking canes.	(-0.33, 0, 10.71)
	9	Grandfather clock	(-2.85, 1.63, 5.853)
	10	Gramophone	(-1.45, 5.548, 4.257)
	11.1	Painting	(8.6, 5.985, 1.524)
	11.2	Painting	(-8.59, 5.985, 1.575)
	11.3	Painting	(0.736, 1.938, 10.96)
	12	Owl	(multiple spawn positions)

compare seated positions within the IVE regarding audio quality or preference. Given the inclusion of human avatars, socio-behavioral task dependencies can be included to study subjects' preferences and immersion. The user task provided for the MANSION IVE is a counting task supported via auditory localization. Counting tasks can be used in cognitive sciences to increase task demand and asses saliency in complex environments. While exploring the mansion, certain elements will visually change and trigger a short audio cue. However, these elements will only change while *not* in the subjects' visual field of view. Thus, subjects rely on the semantic audio content and localization cues to identify (and count) the changed elements. Consequently, audio localization may play an important role in task performance, influenced by factors such as head-related transfer functions, early reflections, and source directivity, for example.

Although the IVEs have been designed with the tasks mentioned above in mind, all scenes are provided open-source, allowing the possibility to extend or alter the IVEs to custom evaluations or alternate tasks. For example, the Mansion scene includes multiple interactive objects, such as chess pieces, bottles, umbrellas, and walking canes, that may be incorporated into a timed collection task. Overall, the three IVEs offer both the possibility for classic audio quality evaluations or alternate avenues for behavioral, cognitive, and quality of experience audio research and development more akin to gaming.

3 Case Study

3.1 Method

The IVEs have been designed to promote various exploration behaviors and accentuate different attributes relevant to acoustic auralization that may impact auditory perception and cognition. To this end, we perform a small-scale case study to demonstrate how these scenes may be used to yield different exploration and to gain insight into audio rendering requirements. To highlight exploration behavior, we record user position data at a sample rate of 0.2 ms throughout the evaluation. For data on audio properties, a questionnaire was employed to gain insight into acoustic attributes that expert users perceive to be relevant for the acoustic auralization of this scene and a given task. Here, an 11-point response scale was given to provide ratings ranging from "not relevant" (0) to "highly relevant" (10). The questionnaire employed is based on previous literature [24], and several acoustic auralization attributes previously identified by a demographic of audio evaluation experts [7], shown in Table 2.

Table 2: Description of how acoustic attributes can be relevant for audio rendering in a particular scene and user task.

Attribute	Scene relevance
Acoustic transitions	If the scene has multiple spaces with different acoustic treatments that can be traversed (i.e., transitioned between) by users.
Diffraction	If the scene has surfaces with reflex angles that require the sound waves to "bend" around.
Early Reflections	If reflections would notably change the perceived audio due to user and source position with respect to the surface.
Late Reverberation	If larger volumes with reflective surfaces resulting in reverberant spaces with persisting audio are present.
Localizability	If it is important that the audio be correctly aligned with a visual source, or the correct position of a sound must be perceived to complete a task.
Near field rendering	If the user can approach sound sources within 1.0 m.
Occlusion	If the direct line of sight between audio sources and the user position can be blocked by surfaces.
Source Directivity	If sources within the scene are more likely to propagate audio in a non-omnidirectional pattern across the frequency spectrum.
Source Extent	If sources are present that have a width, height, and depth greater than point sources.

The case study also allowed to test the performance of the scene on the employed machine. The machine used was a PC with Windows 64-bit 10 OS, with an AMD Ryzen 7 5800X 8-core CPU at 3.80 GHz, 32 GB of memory, and a NVIDIA GeForce RTX 3080 graphics card. The VR hardware employed was the Valve Index and Valve Index controllers (knuckles), and audio was provided over Beyerdynamic DT-770 Pro closed headphones connected via an RME Babyface interface.

The task for all subjects was to explore the scenes as freely as possible and provide ratings for the acoustic attributes. The user tasks described in Sec 2.4 were given to orientate and motivate subjects in their exploration. For the CAVE a way-finding task was used, to explore the cave and try and find the walkie-talkie. For the CINEMA, a preference task was employed, where subjects were to explore the audio rendering at the different seating positions to find their preferred position. For the MANSION, a localization task was employed where subjects must identify (based on the audio cues) objects that change position or placement throughout their exploration. A user interface was provided in VR that subjects could use to rate the relevance of the selected acoustic attributes. The user interface could be toggled on and off at any time.

3.2 Procedure

Subjects were given an introduction to the scenes over a 2D monitor and an explanation of the scene-specific tasks. Afterward, the acoustic attributes were discussed to ensure subjects shared a common understanding. Subjects were then shown the VR controls which included teleportation, UI control, and object interaction. The CINEMA scene provided an optional VR mechanic allowing subjects to teleport to multiple chairs. Before the test started, subjects could practice all interactions in a starting scene until they were comfortable to reduce diluting exploration data with learning behaviors of the VR controls. Finally, the possibility to pause between the three scenes was given as a break-room scene where subjects could remove the VR equipment if needed and resume at any time. No time limit was imposed on subjects, and even after fulfilling any user task, subjects were free to explore at will until deciding to progress. Overall, 14 subjects participated in the evaluation, all employees within Fraunhofer IIS. Of the 14 subjects, 5 were audio experts and were considered in the acoustic attribute scaling.

3.3 Results

Throughout the subject's exploration, the technical performance of Unity and SteamVR was recorded. The starting scene with minimal graphical components was used as a benchmark. This yielded a consistent 80 frames-per-second (fps) observed through the Unity stats window and 5.2 / 12.5 ms frame timing (i.e., the amount of time needed to render the frame). A frame timing above 12.5 ms will start to induce noticeable frame drops. The CAVE scene ran at a mean (M) and standard deviation (SD) of 62 fps and 5.3 ms, respectively. Frame calls were M = 11 ms, SD = 2.5 ms. The CINEMA scene ran at M = 80 fps (SD = 1.2 fps), with frame calls consistently at 7.2 ms. The MANSION scene ran at M = 79.3 fps, SD = 1.3 fps with frame draws of M = 9.4, SD = 1.7.

For the acoustic attributes questionnaire, normalized ratings are shown as acoustic profiles for the three IVEs in Fig. 4. The behavioral data captured throughout subject exploration is illustrated in Fig. 5 using raw position information (colored) and a 2-dimensional

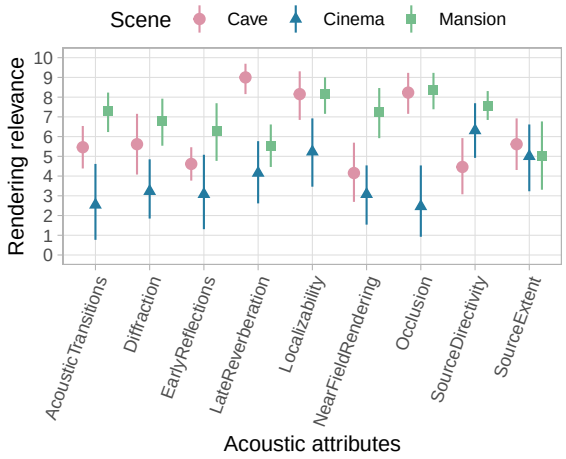

Fig. 4: Mean and bootstrapped 95% confidence intervals for relevance of audio rendering attributes (Table 2) prescribed for the three VR IVEs (0 = not relevant, 10 = highly relevant), while performing the scene task.

(2D) density distribution of the sample data. The 2D distribution represents the frequency (i.e., time spent) throughout the space, thus highlighting areas of most interest to subjects.

3.4 Discussion

The acoustic attributes rated by audio evaluation experts highlight several differences across the IVEs; see Fig. 4. For the CAVE (red), the three attributes rated the highest for audio rendering were late reverberation, localizability, and occlusion. That late reverberation was rated highly is expected, given that most of the scene's materials are mainly dense rock with low absorption coefficients. Localizability was deemed equally relevant as occlusion. Here, not only the physical properties of the space but also the way-finding task may have contributed to these ratings. Objects in the cave, such as a large continuous water stream (Fig. 3 source ID 6), suggest that source extent would be relevant for audio rendering. However, this played no role in the way-finding task and could be why this is rated with lower relevance. For the MANSION scene, localizability was also rated equally high as in the CAVE. The same rating is true for occlusion, even though the physical geometry arguably offers less in occluding surfaces. Given the higher number of sources in the MANSION scene, experts also rated source directivity has highly relevant. While both the CAVE and MANSION tasks are oriented around finding audio sources, the increased relevance of source directivity for the MANSION may be a result of the localization aspect, rather than nagivation. For the CINEMA scene, the relevance for many attributes resulted in lower values. This may be due to the simpler geometric space the intended acoustic treatment of a cinema not evoking many of the specific audio rendering attributes listed in Table 2, as well as a less objective, preference-orientated task. The implication here is that although audio will be affected in the IVE based on the presence of acoustic auralization features, they have little bearing on subjects' ability to complete the task. Importantly, this is not to say that the different rendering features will affect the preference, but simply that subjects can always formulate preference irrespective of the audio rendering. Conversely, audio rendering is objectively tied to the task in the CAVE and MANSION. Consequently, future research and 6-DoF audio rendering may place higher rendering requirements not only on acoustic requirements, but also considering user tasks.

The design of the IVEs also provides differences regarding the exploration behavior, shown in Fig. 5. Due to the level design of the CAVE, subject exploration is primarily a choice between forwards, backward, or intersections, including left and right. The higher density distributions at coordinates (-10, -20) and (-7, -5) also coincide with the spawn positions of the walkie-talkie way-finding objective. For the CINEMA, subjects were also free to explore by teleporting but often chose to switch between selected seats, as shown by the higher density points. Finally, the MANSION scene shows a much higher degree of dispersion throughout the scene, coherent with many of the audiovisual objects throughout the rooms (see Fig. 3). However, exploration is much less 'directed' when compared to the CAVE scene due to the open-plan level design. Consequently, raw sample points illustrated in Fig. 5 are highly scattered between higher-density positions.

The results of this small-scale case study highlight that each scene places a different emphasis on relevant acoustic attributes, as rated by a small group of experts. Moreover, the behavioral data of all users provide a first glance into the impact of the task and level design on user exploration. When conducting sensory evaluations, the choice of stimuli can highly influence results. For audio quality evaluation, multiple genres, including speech and even noise signals, are often selected [25]. Here, the designed IVEs show a diverse set of acoustic attributes and signals relevant

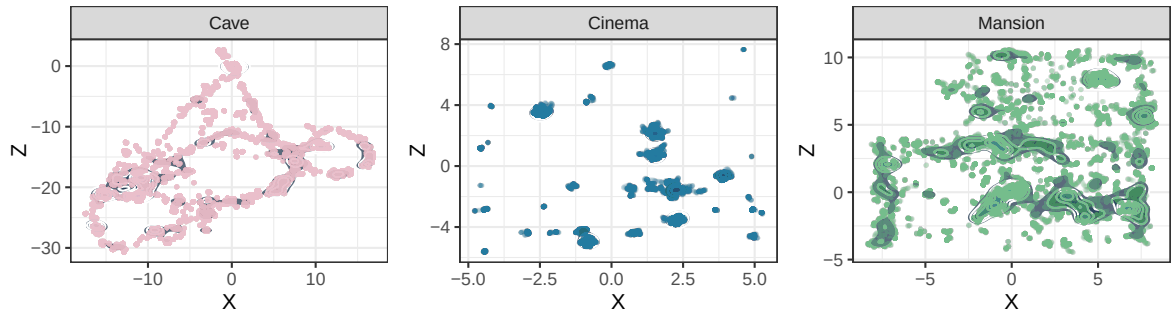

Fig. 5: 2D (x, z) positional exploration data of all subjects across the three scenes expressed as sample points and 2D density plots (left = CAVE, middle = CINEMA, right = MANSION).

for auditory research in VR. Regarding constructs such as QoE, many complex models demonstrate the impact of bottom-up multi-modal sensory input as a feedback loop to further exploration [26]. As highlighted by the case study, each scene offers various exploration styles encouraged by diverse multi-modal sensory information. In combination with interactive VR objects, the designed scenes offer further avenues for QoE research concerning indirect or unobtrusive assessment simply by way of actions and behavior, particularly when given a natural task representative of typical usage [27].

4 Summary

This article describes three VR IVEs designed for audio research and development. The design philosophy of the IVEs considers both audiovisual and interactive cues typical of VR scenarios such as gaming, in addition to game-like task design, to accommodate for more cognitive and behavioral assessment when employing evaluation constructs such as QoE, plausibility, and immersion. The audiovisual design for the IVEs is described, along with the implementation of interactive audio for VR using open-access SDKs and modeled acoustic geometry for more complex acoustic auralization. The design considerations and description of the IVEs are accompanied by a small-scale case study, which demonstrates the impact the scene design has on exploration behavior and provides insights into the relevance of acoustic attributes as rated by audio evaluation experts during a user task. In addition to raw data, further analysis, and resources, the VR IVEs are and available to download at https://www.qoevave.github.io/database/ as individual Unity packages or a combined Unity project.

5 Acknowledgments

The authors would like to thank all Fraunhofer IIS employees who participated in the subjective experiments. This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) priority program SPP 2236 AUDICTIVE, under Project No.: 444832250.

References

[1] Par, S. v. d., Ewert, S. D., Hladek, L., Kirsch, C., Schütze, J., Llorca-Bofí, J., Grimm, G., Hendrikse, M. M., Kollmeier, B., and Seeber, B. U., "Auditory-visual scenes for hearing research," *Acta Acustica*, 6(55), pp. 1–14, 2022, doi:10.1051/aacus/2022032.

[2] Brinkmann, F., Gamper, H., Raghuvanshi, N., and Tashev, I., "Towards encoding perceptually salient reflections for parametric spatial audio encoding," in *AES 148th Convention*, pp. 1–11, Online, 2020.

[3] Terentiv, L., Fersch, C., Fischer, D., and Setiawan, P., "Voxel-based occlusion and diffraction modeling for the upcoming MPEG standard for VR and AR," in *AES Conference on Audio for Virtual and Augmented Reality*, pp. 1 – 10, Redmond, WA, USA, 2022.

[4] Cowan, B. and Kapralos, B., "GPU-based acoustical diffraction modeling for complex virtual reality and gaming environments," in *AES International Conference on Audio for Games*, pp. 1–7, Oshawa, Ontario, Canada, 2011.

[5] Groß-Vogt, K., Höldrich, R., Shen, J., and Duraiswami, R., "Data-driven feedback delay network construction for real-time virtual room acoustics," in *Proc. of the 15th International Audio Mostly Conference*, pp. 46–52, Graz, Austria, 2020, doi:10.1145/3411109.3411145.

[6] Llorca-Bofí, J. and Vorländer, M., "Multi-detailed 3D architectural framework for sound perception research in virtual reality," *Frontiers in Built Environment*, 7(687237), pp. 1–14, 2021, doi:10.3389/fbuil.2021.687237.

[7] Herre, J. and Disch, S., "MPEG-I immersive Audio – Reference model for the virtual/augmented reality audio standard," *J. of the Audio Engineering Society*, 71(5), pp. 229–240, 2023, doi: 10.17743/jaes.2022.0074.

[8] Beckhaus, S. and Lindeman, R. W., "Experiential fidelity: Leveraging the mind to improve the VR experience," pp. 39–50, SpringerWeinNewYork, Heidelberg, Germany, 2011.

[9] Lindeman, R. W. and Beckhaus, S., "Crafting memorable VR experiences using experiential fidelity," in *Proc. of the 16th ACM Symposium on Virtual Reality Software and Technology*, Kyoto, Japan, 2009, doi:10.1145/1643928.1643970.

[10] Pillai, J. S. and Verma, M., "Grammar of VR storytelling: Narrative immersion and experiential fidelity in VR cinema," in *Proc. of the 17th International Conference on Virtual-Reality Continuum and its Applications in Industry (VRCAI)*, pp. 1–6, Brisbane, QLD, Australia, 2019, doi:10.1145/3359997.3365680.

[11] Potter, T., Cvetković, Z., and Sena, E. D., "On the relative importance of visual and spatial audio rendering on VR immersion," *Frontiers in Signal Processing*, 2(904866), pp. 1–10, 2022, doi:10.3389/frsip.2022.904866.

[12] Neidhardt, A. and Zerlik, A. M., "The availability of a hidden real reference affects the plausibility of position-dynamic auditory AR," *Frontiers in Virtual Reality*, 2(678875), pp. 1–17, 2021, doi:10.3389/frvir.2021.678875.

[13] Letter, J. D., Zheleva, A., Maes, M., All, A., Marez, L. D., and Durnez, W., "What did you expect? Modelling quality of experience for virtual reality using the repertery grid technique," *Quality and User Experience*, 6(5), pp. 1–13, 2021, doi:10.1007/s41233-021-00045-6.

[14] Hofer, M., Hartmann, T., Eden, A., Ratan, R., and Hahn, L., "The role of plausibility in the experience of spatial presence in virtual environments," *Frontiers in Virtual Reality*, 1(2), pp. 1–9, 2020, doi:10.3389/frvir.2020.00002.

[15] Cutler, L., Darnell, E., Dirksen, N., Hutchinson, M., and Peterson, S., "Crow: The Legend," in *Proc. of the ACM SIGGRAPH 2018 Virtual, Augmented, and Mixed Reality*, pp. 1–1, Vancouver, British Columbia, Canada, 2018, doi:10.1145/3226552.3226578.

[16] Poirier-Quinot, D. and Katz, B. F. G., "Impact of a HRTF individualization on player performance in a VR shooter game II," in *AES Conference on Audio for Virtual and Augmented Reality*, pp. 1–10, Redmond, WA, USA, 2018.

[17] Rummukainen, O., Schlecht, S. J., Plinge, A., and Habets, E. A. P., "Evaluating binaural reproduction systems from behavioral patterns in a virtual reality – A case study with impaired binaural cues and tracking latency," in *AES 143rd Convention*, pp. 1–8, 2017.

[18] Garí, S. V. A., Calamia, P., and Robinson, P., "Navigation of virtual mazes using acoustic cues," in *AES 154th Convention*, pp. 1–10, Helsinki, Finland, 2023.

[19] Popp, C. and Murphy, D. T., "Establishment and implementation of guidelines for narrative audio-based room-scale virtual reality using practice-based methods," in *AES Conference on Audio for Virtual and Augmented Reality*, pp. 1–10, Redmon, WA, USA, 2022.

[20] Robert, F. A. S., Wu, H.-Y., Sassatelli, L., Ramanoel, S., Gros, A., and Winckler, M., "An integrated framework for understanding multimodal embodied experiences in interactive virtual reality," in *ACM International Conference on Interactive Media Experiences (IMX)*, pp. 1–14, Nantes, France, 2023.

[21] Lee, B., Rudzki, T., Skoglund, J., and Kearney, G., "Context-based evaluation of the opus audio codec for spatial audio content in virtual reality," *J. of the Audio Engineering Society*, 71(4), pp. 145–154, 2023, doi:10.17743/jaes.2022.0068.

[22] Paterson, J. and Kadel, O., "Audio for extended realities: A case study informed exposition," *Convergence: The International Journal of Research into New Media Technologies*, 0, 2023, doi:10.1177/13548565231169723.

[23] Delerue, O. and Warusfel, O., "Authoring of virtual sound scenes in the context of the listen project," in *AES 22nd International Conference on Virtual, Synthetic and Entertainment Audio*, pp. 1–9, Espoo, Finland, 2002.

[24] Lindau, A., Erbes, V., Lepa, S., Maempel, H.-J., Brinkman, F., and Weinzierl, S., "A Spatial Audio Quality Inventory (SAQI)," *Acta Acustica united with Acustica*, 100(5), pp. 984–994, 2014, doi:10.3813/aaa.918778.

[25] Bech, S. and Zacharov, N., *Perceptual Audio Evaluation - Theory, Method and Application*, John Wiley & Sons, Chichester UK, 2006, doi:10.1002/9780470869253.

[26] Raake, A. and Blauert, J., "Comprehensive modeling of the formation process of sound-quality," *International Workshop on Quality of Multimedia Experience (QoMEX)*, pp. 76–81, 2013, doi:10.1109/qomex.2013.6603214.

[27] Raake, A. and Wierstorf, H., "The Technology of Binaural Understanding," *Modern Acoustics and Signal Processing*, pp. 393–434, 2020, doi:10.1007/978-3-030-00386-9_14.